Data Service Infrastructure for the Social Sciences and Humanities

EC FP7

Grant Agreement Number: 283646

DASISH Work Package 6: "Legal and Ethical Issues"

Deliverable: D6.1
Deliverable Name: "Report about New IPR Challenges"
Deadline: 30th June 2013
Nature: R

Responsible: MEA (MPG), CITY
Work Package Leader: MEA (MPG)

Contributing Partners and Editors: Daniel Schmidutz, MEA (MPG)
Lorna Ryan, CITY
Anje Müller Gjesdal & Koenraad De Smedt, UiB

Abstract

Research ethics and the legislative regime(s) within which social sciences and humanities (SSH) research takes place have long been the concern of researchers within the SSH domain. Technological advances giving rise to new methods of data collection and data creation amplify the ethics issues which need to be addressed and extend the requirements for researchers in relation to legal provisions. Allied with these technological changes, cross-national research raises particular ethical and legal challenges to which researchers have to respond and researchers' codes of practice may not have been developed apace.

After introducing into the content and presenting the objectives and methodology of the DASISH work package 6 (Sections 1 and 2), the *Report about New IPR Challenges: Identifying Ethics and Legal Challenges of SSH Research* opens with an overview of key principles guiding research ethics in the social sciences and humanities (Section 3.1).

In the following chapters, the different disciplinary and methods specific codes of ethics and professional practice are presented with particular attention paid to cross-national research ethical practice (Section 3).

Subsequently, the approach to ethics review in the EU Framework Programme 7 is presented and as part of the administrative framework of transnational research some aspects of the organisation of the ethics committee review systems across the EU Member States, with a special focus on cross-disciplinary research including the collection of biological samples, are presented (Section 4).

The legal context of the operation of the SSH ESFRI research infrastructures (RIs) and of individual researchers is currently characterised by disquiet in the face of proposed amendments to the anticipated European General Data Protection Regulation. Besides the legal regime relating to data protection and data privacy, intellectual property rights relating to copyright and database rights are relevant to the SSH RIs (Section 5). The specific legal issues faced by researchers may require direct legal input and so a generalist rather than specialist discussion about these issues is presented.

This general discussion is followed by a presentation of how new ethical and legal issues in the SSH domains can be identified and differentiated, including an overview of key issues (Section 6). Subsequently, some specific ethical and legal matters relating to informed consent, bio-social research, harvesting language data, data processing (including data linkage) and data access and data re-use are presented (Section 7).

The report concludes (Section 8) by considering the implications of the proposed General Data Protection Regulation. There is a consensus that this will have severe ramifications for research in the social sciences and humanities domains. Furthermore, the on-going development of and reflection on ethics codes for interdisciplinary, cross-national research remains a challenge for professional associations and RIs.

Finally the most important **results of the report** with regard to present and future legal and ethical challenges of SSH research are summarised (Section 8.4).

Report about New IPR Challenges:
Identifying Ethics and Legal Challenges of SSH Research

CONTENT

1 Introduction

This report of work package 6 "Legal and Ethical Issues" of the "Data Service Infrastructure for the Social Sciences and Humanities" (DASISH) project explores the extent and nature of IPR challenges confronting researchers in the social sciences and humanities (SSH) domains, with special focus on novel challenges which have arisen.

The report concentrates on the identification of new **ethical issues** and **legal requirements** occurring in the SSH domains in relation to 'new' data types or connected to new ways of data collection and linking research data with data from external sources. It represents a first step in order to support researchers in the SSH to cope with new ethical and legal challenges in that regard and to provide practical assistance to the SSH ESFRI research infrastructures (CESSDA, CLARIN, DARIAH, ESS, SHARE) as well as the related research communities in their day-to-day operations of SSH data collection, curation and dissemination.

Since existing more 'traditional' data sets in the social sciences mainly consisting of self-reported information of interviewees, for example, are limited, modern research requires the collection of much more innovative variables, such as objective health information or precise data on the financial status of the respondents. Regarding this, for instance, the inclusion of biomarkers (i.e. objective measures of biological functions) and the linkage of survey data with administrative records recently gained in importance for field surveys. Furthermore, paradata generated in the process of survey production, some of which may be of a sensitive nature and new data sources – especially the Internet with regard to social media data, written and spoken language data (fragments/texts) accessible via websites, for example – have become increasingly important for research in the humanities and social sciences. The extent to which these types of data and ways of collecting and linking data impose new legal and ethical challenges for the SSH, including new and special data protection requirements requires attention, not least due to the legal and ethics issues and associated procedures to which they give rise in the day-to-day operation of SSH data collection.

Focussing on issues that are currently concerning the five SSH ESFRI research infrastructures (RIs) participating in the DASISH project (CESSDA, CLARIN, DARIAH, ESS, SHARE), the objective of this report is to provide an overview of central ethics issues and legal constraints and requirements related to the collection of 'new' types of data being recorded in the SSH domains. In this connection, the report aims to identify new challenges arising in cross-country research that require attention over the course of the research and data curation and data dissemination lifecycles.

In this respect it is noted that the legislative regime which impacts upon the governance of the research process (including ethics), from research project/study initiation to data

curation, is marked by fragmentation and uncertainty – the current negotiations relating to the EC Proposal of a "General Data Protection Regulation" will impact on the data protection regime currently in place. The legal basis of this regime is the "Data Protection Directive" (95/46/EC) and associated national data privacy regulations/laws (i.e. the various implementations of the Directive 95/46/EC). This is further elaborated in Section 5 and Section 8.3. Against this background, this report also aims to indicate the legal context and key elements of ethics guidelines and frameworks that have to be taken into account in relation to SSH data collection, curation and dissemination.

2 Objectives and Methodology

2.1 Overall Objectives of Work Package 6

Work Package 6 (WP6) of the DASISH project addresses various legal and ethical issues that modern research in the social sciences and the humanities (SSH) is confronted with. The focus of WP6 is on legal and ethical issues for data collection, data curation and preservation and data dissemination in the SSH domains. Following the "Description of Work" (DoW), Annex 1 to the Grant Agreement of the DASISH project, WP6 has the following main objectives:

- to identify the legal and ethical issues, constraints and requirements for all data types occurring in the SSH domain as result of data integration and linking;
- to deal with the legal and ethical challenges imposed by the new data types emerging in the social sciences and humanities;
- to look for professional long-run preservation strategies and policy-rules that can be applied to data collections in the SSH.

2.2 Objectives of Task 6.1 and Focus of Deliverable D6.1

Task 6.1 of WP6 "New Ethical and Legal Challenges" is intended to serve the overall objectives of WP6, focussing on 'new' data types being recorded in modern research in the social sciences and humanities. In accordance with the DoW the specific objectives of Task 6.1 are

- to identify the various new data types including sensitive data;
- to determine the IPR Requirements for these new data types in a multi-country usage scenario;
- to define guidelines for appropriate measures and identify tools that help to take appropriate measures.

In this respect it is noted that the DoW presumed that 'new data types' existed and, as a concomitant, that 'new IPR requirements' exist. However, looking at the entire description of Task 6.1 in the DoW, it becomes apparent that the data types referred to, such as biomarkers[1], paradata[2] or process generated data (i.e. administrative record data) are not essentially 'new' types of data. Rather, it appears that the mode or the extent of the collection or linkage of these data and/or the research context, including the intended use of the data, are new within the SSH domains.[3] Nevertheless, since the experience for researchers in the SSH domain is new, specific types of data, including legal requirements and ethical issues related to them (and in some cases maybe even because of these aspects), are perceived as something new.

In particular due to the lack of experience of researchers in the SSH domains with regard to the collection and handling of these data that have become increasingly important in modern research, there is a need to identify the various legal requirements and ethical issues related to the collection and linkage of these data types. Guidelines for appropriate data protection measures have to be defined and appropriate procedures for access have to be investigated in order to deal with legal and ethical challenges resulting from the integration of these data and the linking with data from external sources. Such guidance, of course, has to be informed by the legal provisions as set out in existing and emerging laws and regulations.

This *Report about New IPR Challenges* – as the main output of Task 6.1 of WP6 – therefore focuses on the identification of ***ethics issues*** and ***legal requirements*** related to those data types, which are perceived as something new and/or associated with new data collection and linking practices in the SSH domains. It can be considered as a first step to support SSH researchers as well as the SSH ESFRI RIs to cope with new ethical and legal challenges in that regard and to provide practical assistance to the related research communities with regard to the day-to-day operations of SSH data collection, curation and dissemination.

Adopting Bainbridge's definition of 'intellectual property law' as "the area of law which concerns legal rights associated with creative effort or commercial reputation and goodwill" (2007: 3) and noting that key areas of intellectual property law relate to patents, design right, trademarks and copyright, this report assumes that *copyright* is the area in intellectual

[1] Biomarkers are derived from body fluids and therefore include the collection of biological material, such as blood samples or saliva.

[2] Paradata can be defined as micro-level data about the process of survey production, including data recorded as a by-product in the course of conducting a survey (such as listing information, keystroke data, contact data and gross sample data) as well as auxiliary paradata, i.e., additional data obtained separately from external sources or with a specifically targeted effort to enhance the information on the survey production process (such as interviewer observations, interviewer demographic characteristics, external supplementary data about the sample cases).

[3] For example, biomarkers, such as glycated hemoglobin (HbA1c), a marker of diabetes, have been collected for many years in clinical trials and medical/epidemiological research already – consequently, collecting biomarkers in the course of a population-based field survey does not make them a "new" data type.

property law that is of most relevance to social science and humanities researchers[4]. Besides, for SSH data collectors as well as data archives *database rights* are of relevance (cf. Section 5.2).

The concept of intellectual property rights (IPR), however, merely covers a very small part of the legal and ethical aspects that are of relevance with regard to the different phases of data collection, data processing and data curation in SSH research. In addition, another area of law, specifically that relating to *data protection* and data re-use is also relevant. This area of law covers for example, data privacy and data transfer arrangements between data collectors, archives/data facilities and data users.

In this respect it is noted that even though the description of Task 6.1 in the DoW already clearly suggests that – besides IPR – other legal and ethical aspects, in particular issues related to confidentiality and data privacy, have to be taken into account as well, the original title of the report only refers to IPR. Insofar it was regarded as necessary to expand the initial focus of this *Report on IPR Challenges* accordingly (as is indicated by the subtitle: *Identifying Ethics and Legal Challenges of SSH Research*) in order to serve the overall objectives of WP6 in an appropriate way. As the report will show, modern research in the SSH is confronted with various other (new) legal and ethical challenges in the course of data generation and management that go far beyond the notion of IPR (even though these challenges may lead to certain IPR-related issues).

In the context of WP6 of the DASISH project, the outcome of D6.1 will feed into Task 6.2 ("Virtual L&E Competence Centre") and D6.5 ("Handbook on legal and ethical issues for SSH data in Europe").

2.3 Methodology and Procedure

In the initial phase of cooperation in WP6, specific attention has been paid to the improvement of the common understanding of the day-to-day legal and ethical challenges that occur in the different research infrastructures participating in the DASISH project (CESSDA, CLARIN, DARIAH, ESS, SHARE). To gain deeper insights in each other's activities seemed to be crucial in order to identify distinct legal and ethical aspects that are of common interest for (two or more of) the involved partners as well as to identify differences in this regard.

The discussions and reflections in this context, have been particularly helpful regarding the next step: the identification of the various 'new' data types (including sensitive data) in the

[4] It is noted that copyright is a specific form of right subsisting in various works. "The author of a work is the person who creates it and he […] is normally the first owner. Copyright gives the owner the right to do certain things in relation to the work which includes making a copy […]. Ownership of a copyright […] can be transferred to another or a licence may be granted by the owner to another, permitting him to do one or more specified acts with the work in question" (Bainbridge, 2007: 5).

SSH. Besides the data types mentioned in the DoW (such as biomarkers, paradata and integrated administrative record data), which are of particular importance in the context of survey research, further types of data that increasingly gain importance in the humanities (such as social media data, previously published language data or audio-visual data) have been identified. As part of this task the nature and extent of different types of data occurring in the SSH domain have been discussed.

Subsequently, the legal constraints and requirements (as set out in national and international data protection legislation) and ethical issues (e.g. in connection with ethics committee approvals) for the different data types have been investigated and the major legal and ethical challenges of SSH research related to these data types, focussing on the collection of biomarkers, the collection and use of language data and on the linkage of survey data with administrative records in a multi-country setting, have been compiled and discussed. In this context, a differentiation of legal and ethical challenges according to the different stages of the research process, including data generation and data management has been applied (see Section 6).

The work in Task 6.1 can be illustrated by two examples: On the one hand, for example, the ethical aspects and the legal requirements of different types of paradata, which are collected in the context of survey research, have been investigated systematically. The major legal and ethical challenges connected to paradata collection and use are outlined in Section 7.3.2 (please also see deliverable D6.2 of the DASISH project for a more detailed presentation of the respective investigations). On the other hand a transnational systematic inquiry regarding national legal requirements and ethics committee approval procedures in the EU with regard to the collection of biomarkers (derived from dried blood spots) has been carried out by MEA (MPG) making use of the SHARE Research-Network (see Annex 10.4). First findings of this inquiry are presented in Section 7.1.2 of this report.

3 Ethics in the Social Sciences and Humanities

According to Denscombe, ***ethics*** concern what *'ought'* to be done (2002: 175); they are "a matter of principled sensitivity to the rights of others" (Bulmer, 2001, p. 46). However and beyond that, ethics are linked to matters of professional integrity (cf. Denscombe, 2002) and related to the wider reputation of social science and humanities research:

> "Deception, manipulation and abuses of trust need to be avoided both as morally harmful but also because they can provoke a backlash against social science" (Bulmer and Warwick, 1993: 19).

In line with this, Flew suggests that ethics are

> "a set of standards by which a particular group or community decides to regulate its behaviour – to distinguish what is legitimate or acceptable in pursuit of their aims from what is not" (1979: 112).

However, ethics are not officially sanctioned codified rules. In the realm of ethics researchers' positions become relevant and, as Iphofen (2011) notes, ethical pluralism, involving 'normative ethics', poses a challenge for researchers.

3.1 Key Principles of Research Ethics

Ethics, Denscombe states (2002: 176), call for a moral perspective rather than a practical perspective. Research ethics seek to protect the interests of research participants and are guided by the core principles of

- 'do no harm',
- informed consent,
- protection of anonymity and
- confidentiality.

It is the responsibility of the researcher and the responsible data handling body to ensure that the participants of the research are not harmed by their participation in research. It requires that the researchers and others, such as archiving bodies, consider and, as appropriate, communicate to those participating in the research, the intended and unintended consequences of the research.

Denscombe (2002) identifies "consent, authorisation, integrity and data security" as central to the recognition of the rights and interests of participants in social research. He suggests that one of the ground rules for social research requires that "researchers need to recognise the rights and interests of participants" (ibid.). This may be seen to have occurred when

> "[d]ue consideration has been given to the impact of the research on those affected by it, and where it has been reasonable to do so, informed consent has been obtained from those directly involved in the research. Where appropriate, measures have been taken to maintain the confidentiality of information and minimise intrusion into people's lives." (Denscombe, 2002: 174)

In this connection it is noted that the key development in considerations of ethics relates to the Internet and related technological advancements (such as innovations in data mining techniques), allowing for the collection of vast amounts of data from numerous sources[5] as well as providing a platform for data collection (e.g. in form of online surveys).

[5] I.e., which allow for a data harvesting capability that is termed the 'data deluge' (UK Data Forum 2009: 17).

Technological developments have implications for how the ethical and legal aspects of the entire research process are e considered. While it is generally accepted that the online environment *amplifies*, rather than transforms, the nature of research ethics and legal requirements relating to research, it is also clear that these developments give rise to specific issues which are not covered in existing ethics codes. For example, web mining, the process of gathering data from a range of Internet sources – also referred to as 'invisible information gathering' (Van Wel and Royakkers, 2002: 129) – poses particular challenges with respect to informed consent; and the same holds for text data mining, a technique to discover new (i.e. previously unknown) information by automatically extracting information from textual databases. "[T]he ethics issues may not be new but the new technologies make it possible to mine data in new ways" (ibid.: 134), which by their different nature in comparison to 'traditional' data collection techniques amplify exactly those ethics issues.

3.2 Ethical Guidelines and Codes of Ethics

With regard to ethical guidelines, Denscombe (2002) comments, that there is no shortage of guidance when it comes to ethics. The codes vary a little but generally key themes can be identified which are common to all. He notes that codes do not constitute rules:

> "The point is not that each principle should be *followed*, but that it should be taken into account and *considered.* Each principle provides a starting point, a baseline against which to compare the actual position adopted by the researcher. If circumstances arise where the researcher feels that he or she is not able to be bound by a specific principle it becomes necessary to weigh the pros and cons of the situation and to arrive at a decision about whether it is legitimate to 'relax the rules' on this occasion. To do so does not automatically condemn the research as 'unethical' but it does warrant some explanation. *The principle should be acknowledged*." (Denscombe, 2002: 176)

Following Bulmer (2001), the position advanced by 'virtue ethics' (cf. Kimmel, 1988) is that researchers and others with responsibility for data should demonstrate sensitivity to the different ethics issues raised during all phases of the research process.

Annex 10.2 lists some, not all, of the existing codes of ethics and associated guidance on good professional conduct and research integrity and the following table provides an overview of how these might be categorised.

Ethics	International	EU	National (e.g.)
SSH Disciplines	International Sociological Association	RESPECT *Code of Practice*:	British Sociological Association (BSA): http://www.britsoc.co.uk/m

(e.g. Sociology)	(ISA): http://www.isa-sociology.org/about/isa_code_of_ethics.htm	http://www.respectproject.org/code/	edia/27107/StatementofEthicalPractice.pdf German Sociological Association (GSE) & Berufsverband Deutscher Soziologen (BDS): http://www.soziologie.de/index.php?id=19
Funding bodies	Management of Social Transformations (MOST, an international programme established by UNESCO): http://www.unesco.org/most/ethical.htm	RESPECT *Code of Practice*: http://www.respectproject.org/code/ Framework Programme 7: http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=129	ESRC *Framework for Research Ethics*. 2010. ESRC London: http://www.esrc.ac.uk/about-esrc/information/research-ethics.aspx,
Statistics	International Statistical Institute: http://www.isi-web.org/images/about/Declaration-EN2010.pdf	European Statistics *Code of Practice*: http://epp.eurostat.ec.europa.eu/portal/page/portal/quality/code_of_practice	National Statistical Agencies; e.g. National Statistics *Code of Practice* (UK): http://www.ons.gov.uk/ons/guide-method/the-national-statistics-standard/code-of-practice/national-statistics-code-of-practice/index.html
Practitioners (Market Research)	WAPOR (World Association for Public Opinion Research): http://wapor.unl.edu/wapor-code-of-ethics/	ICC/ESOMAR (European Society for Opinion & Marketing Research) *International Code on Market and Social Research*: http://www.esomar.org/index.php/professional-standards.html	Market Research Society MRS (UK) *Code of Conduct*: https://www.mrs.org.uk/standards/code_of_conduct/

As can be seen in the table and in Annex 10.2 there are various ethical guidelines and codes of ethics available at all sorts of institutional and professional levels. However, with regard to technological developments which lead to new practices, it has to be noted that specific ethical issues may arise which are not covered in existing ethics codes. Researchers and research communities have to respond dynamically to these. Especially concerning the growing importance of Internet-based research (e.g. regarding web surveys or the harvesting of language data from the Internet), Charlesworth' assessment is apposite:

> "The degree to which ethical guidelines/frameworks are directly helpful in addressing e-social science research-related issues tends [...] to relate heavily to the interests of those who have developed them, and the extent to which the ethical guidelines/frameworks have been subject to continuing review and revision." (2012: 93)

The Association of Internet Researchers' Ethics Guide and the British Psychological Society Code of Human Research Ethics provide examples of continuing review. While the Association of Internet Researchers has a dedicated ethics wiki which facilitates continuous discussion and exchange (http://ethics.aoir.org), the British Psychological Society has issued complementary guidance for Internet-based research to its general Code of Ethics (http://www.bps.org.uk/publications/policy-and-guidelines/research-guidelines-policy-documents/research-guidelines-poli).

3.3 Cross-National Research Ethics

The 'new' legal and ethical challenges affecting SSH researchers are not only national in scope; they are international. There is relatively less focus on research ethics and legal issues arising from international and cross-national research. Bulmer and Warwick (1993) address issues relating to international research, and Oyen (1990) raises ethics in considering comparative research methodology.

Cross-national survey guidelines have been published in 2012 as part of the Comparative Survey Design and Implementation (CSDI) Guidelines Initiative (http://ccsg.isr.umich.edu/). The ethics guidelines, presented in Section III of the Cross-Cultural Survey Guidelines (CCSG), note that ethics issues arise and have to be addressed throughout the survey lifecycle

> "[t]o ensure that participating research teams follow widely accepted standards for ethical, professional, and scientific conduct from the design of the study through implementation, dissemination, and reporting." (Alscer et al., 2012: III.4)

In this connection, the ethics guidelines provide an overview of the key elements to be addressed; outlining the rationale for addressing these elements and, importantly for survey researchers and others involved at different points in the survey lifecycle, providing full explication of procedural steps. Crucially, they point out that

"[p]roper, ethical conduct may be simple and straightforward in one location but require multiple steps in another." (Alscer et al., 2012: III.10)

The main issues requiring attention in ensuring ethical research practice are the dual commitment to protect the rights of free will, privacy, confidentiality and well-being of research participants, and minimize the burden of study participation to the greatest extent possible. This requires that researchers avoid undue intrusion and obtain voluntary informed consent. Key principles include:

- "Do not use coercion. Whether a practice is defined as coercive or not may vary by culture, population, and study. Large monetary payments that are given to participants may be considered coercive in some studies.
- Respect the rights of individuals to refuse to be interviewed, to refuse part of the interview, and to terminate an interview in progress. Whether or not follow-up with individuals who initially refuse the survey request is appropriate may vary by culture, population, and study.
- Respect the right of individuals to refuse to answer any question in the interview.
- Obtain and document consent. Whether consent is obtained in oral or written form depends on a number of factors, including government laws and regulations, risk of harm for respondents revealing sensitive information, the mode of data collection, the type of information requested, and cultural norms. In mail surveys, consent may be implied (that is, not explicitly obtained in oral or written form) if the respondent chooses to fill out the questionnaire and mail it back.
- Obtain informed consent from a parent or responsible adult before interviewing children or young people.
- Avoid making inaccurate or overly restrictive statements (e.g., the data will only be shared with the research team) if the data will be archived and shared with the research community
- Consent information should be conveyed in a format that is easy for respondents to understand. […]
- Make clear to respondents the extent to which confidentiality is protected.
- If disclosing survey data to outside parties, require all subcontractors, consultants, and third parties to enter into an agreement to maintain respondent confidentiality. This agreement should include an explicit statement that the outside party cannot use contact information or any other information to re-contact the respondent for any reason not directly related to the study (e.g., data cannot be used to approach respondents for a different study or for marketing purposes)"
  (Alscer et al., 2012: III.7-8)

The authors recommend that full records are kept including:

- "Scripts, letters, fact sheets, and any other materials provided to respondents to give them information they need to make an informed decision about participation.

- Consent form templates and protocols.
- Translated or adapted consent form templates and protocols.
- Individual consent information for each respondent, stored in a safe environment separate from survey data.
- Confidentiality procedures and protocols.
- Pledge(s) of confidentiality completed by staff.
- Records of completion of any specialized staff training on ethics.
- Ethics review board original submission and requests for modification to study protocol (see Appendix D for a checklist of materials to include an ethics review board submission).
- Ethics review board correspondence (e.g., letters of approval).
- Any correspondence between study staff or ethics review board members/staff and respondents regarding an ethical issue or concern.
- Reports of quality control activities (e.g., documentation of verification activities)."
  (Alscer et al., 2012: III.18-19)

4 Institutional and Administrative Frameworks

Although ethics call for a moral perspective rather than a practical perspective (cf. Denscombe, 2002: 176), researchers in the SSH domains have to acknowledge the requirements of the institutional context and the administrative frameworks of their research. With regard to the SSH ESFRI RIs two aspects have to be taken into account in that regard: the ethics review in the EU Framework Programme 7 and – depending on the concrete research context – additional reviews by national or local research ethics committees.

4.1 Ethics Review in Framework Programme 7

For the SSH ESFRI research infrastructures in particular the ethics review of the Seventh Framework Programme for Research and Technological Development (FP7) has to be considered. Ethics review carried out as part of a FP7 project proposal assessment is

> "a legal requirement […] and is intended to ensure that all research activities carried out under the Framework Programme are conducted in accordance with fundamental ethical principles. The Ethics Review evaluates aspects of the design and methodology of the proposed research that raise ethical concerns. These may include intervention on humans, use of animals, data protection issues, use of children, and research proposed to take place in developing countries" (http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=1289&lang=1, accessed 27/06/2013)

Project proposers for FP7 grants are required to identify any ethical issues in their proposals (in Part B, Section 4) and all proposers must complete the 'Ethics Table' provided, which asks about the involvement of human subjects in the research. The Ethical Rules in FP7 main page can be accessed here: http://ec.europa.eu/research/science-society/index.cfm?fuseaction=public.topic&id=129 and full details of the Ethics Review process are presented on the CORDIS pages: http://cordis.europa.eu/fp7/ethics_en.html#ethics_cl.

In this connection also the aforementioned EU-wide RESPECT Code of Practice that has been published in 2004 should be mentioned. It "is based on a synthesis of the contents of a large number of existing professional and ethical codes of practice, together with current legal requirements in the EU. Whilst the RESPECT provisions are voluntary, some of the requirements on which they are based are morally binding on the members of specific professional associations or legally binding on citizens of EU Member States." (RESPECT, 2004)

The RESPECT Code of Practice is based on three main principles:

1. Upholding scientific standards,
2. Compliance with the law,
3. Avoidance of social and personal harm.

More than the codes of ethics of the professional associations, the RESPECT Code gives equal priority to the legal issues and presents an overview of the key intellectual property and data protection concerns, with suggested courses of action. These are reproduced in Annex 10.3.

4.2 Research Ethics Committees Review Systems

Research ethics committees (RECs) operate at the interface between the legal system and ethical frameworks. The scope of committee work is usually not governed by law but takes place within the wider framework of research governance. However, ethics committees often have responsibility to assess proposals with reference to legal provisions, such as in relation to informed consent of different groups, including children and those identified as vulnerable. Signing off on research proposals by a research ethics committee may be linked to indemnifying the researcher against any legal action arising from the research.

The organisation of national research ethics committee systems, however, differs a lot between the different EU Member States. This becomes particularly apparent when conducting cross-disciplinary research that includes the collection of biological samples and

bio-medical research ethics committees become involved.[6] When the conduction of transnational and transregional research projects that require approval of ethics committees from all participating countries are intended, for each country the ethics committee/s responsible has/have to be identified and applied to in accordance with the respective national or regional policies and procedures. Some countries have one single national ethics committee easy to identify. But in other countries several ethics committees have to be involved (e.g. in Italy, Spain or Switzerland). Sometimes, this is only the case if the study is considered to be a 'multicentre study' (as e.g. in Belgium), sometimes, at any rate for every region where respondents live another ethics committee has to be involved (as e.g. in Spain and Italy). The most extreme example of this probably is Italy. In Italy, the ethics committees are organised at the municipality level (in 2007 there were 310 ethics committees) and they are independent, i.e. each committee has its own rules, procedures and schedule. If the research is conducted in more than one municipality, applications have to be made at each municipality's ethics committee separately.

In addition, the requirements with regard to research projects vary a lot in the different countries. And even the requirements respectively imposed restrictions to research projects of different ethics committees within a single country might vary a lot as could be experienced in Belgium in the context of the ethics approval of the dried blood spots collection in SHARE (when one ethics committee approved an application, while another committee did not approve exactly the same application in the first instance[7]). This also might be connected to a different composition and alignment of the different ethics committees. While in some countries there is a differentiated ethics committee system, in which different ethics committees exist for different research purposes (e.g. social sciences and clinical trials), in other countries bio-medical research ethics committees are predominant, which mainly review clinical or medical epidemiological studies (e.g. Poland and Portugal). In the latter case, bio-medical RECs (which might give precedence to bio-medical ethics over SSH ethics requirements and sometimes might overlook the intrinsic differences in substance and methodology between clinical and social science research) also claim to be responsible to review cross-disciplinary studies such as SHARE.

In contrast to that, in the UK or Germany, for example, research ethics committees are not only established in the area of bio-medical research, but also on a university-level. In the UK, for instance, universities have their own RECs. Here, the scope of committee work is clearly not governed by law but takes place within the wider framework of research governance within universities. Social science and humanities researchers, generally working within the

[6] The website of the European Network of Research Ethics Committees (EUREC) provides an overview of Research Ethics Committees in Europe from a bio-medical point of view: http://www.eurecnet.org/index.html.

Another overview from the same perspective can be found on the website of the PRIVIREAL project, which has been examining the implementation of the Data Protection Directive 95/46/EC in relation to medical research as well as the role of ethics committees: http://www.privireal.org/content/rec/countries.php.

[7] However, finally the dried blood spots collection was approved by both ethics committees.

education sector, are more likely to access university RECs. The figure below provides an itemised listing of the key points generally considered by RECs in the UK:

ETHICAL ISSUES COVERED IN A UNVERSITY ETHICS REVIEW FORM (UK)

- Have permissions been sought from appropriate external bodies [...]
- Justification of the scientific method
- Plans for dissemination
- Self-identification of potential ethical issues
- Identification of benefits likely to accrue to participants and society at large
- Details of any physically invasive or psychologically intrusive procedures to be used
- Details of precautions in place to minimize harm to participants
- Details of physical and psychological assessments to determine suitability to be a participant
- Details of how participants are to be recruited.
- Details of safeguards in place ensuring that no coercion or pressure is placed on participants to take part.
- If vulnerable people are to be asked to participate what extra safeguards are in place to protect their wellbeing
- Details of the study inclusion and exclusion criteria; and justification
- The procedures for obtaining informed consent
- Details of the steps in place to refer participants on to further help should a need be identified.
- How will participants be protected from over-research
- How will the health and safety of researchers be established
- What steps are in place to ensure confidentiality and security of data and compliance with the Data Protection Act [UK]
- Details of how informed consent will be secured
- Contact with NRES [National Research Ethical Service in the UK] if clinical dimension to research
- Recruitment
- Confidentiality
- Details of data management and storage arrangements

(Extracted from Ryan, Cooper and Drey, 2013)

5 Legal Context/Framework

In contrast to ethics which concern what 'ought' to be done, ***law*** concerns what *'should'/must* be done. Laws are officially sanctioned codified rules governing behaviours and practices the breach of which incurs sanction. The researchers' (value) positions are not relevant in the legal context, what is relevant is whether the behaviour is in variance with statutory requirements. The law generally sets out what can and cannot be done – ethics

guidelines and frameworks generally set out different positions. Even though there are a lot of commonalities between ethical practice and legally required practice, which are especially evident in considering issues of anonymity and data protection issues, law has primacy over ethics codes.

The legal framework within which the SSH research infrastructures operate is governed by two sets of legislative provisions:

1. Data protection (privacy and autonomy)
2. Copyright & database right (intellectual property)

In particular, the "Data Protection Directive" (95/46/EC) respectively the national implementations of this Directive and the anticipated "General Data Protection Regulation" apply to the first set of provisions and a similar mix of national and European law as well as international conventions govern copyright, which is part of the second set of provisions.

The RESPECT Code of Practice recommends the following in relation to compliance with the law

> "In general, socio-economic researchers should comply with the laws of the countries in which they are based or in which they are carrying out research. In the case of international collaborations or online research, the laws of additional countries may also apply. Researchers have a duty to ensure that their work complies with any relevant legislation. Two areas of law (data protection law and intellectual property law) are particularly relevant for the conduct of research, especially research involving human subjects, and researchers should acquaint themselves with the relevant national and international provisions." (RESPECT, 2004)

At this, European Union codes and law should take *precedence* at particular points – particularly in the transfer of data across national borders. The World Medical Association Declaration of Helsinki, which includes the 'Ethical Principles for Medical Research Involving Human Subjects', serves as a good example on how precedence is often expressed:

> "Physicians should consider the ethical, legal and regulatory norms and standards for research involving human subjects in their own countries as well as applicable international norms and standards. No national or international ethical, legal or regulatory requirement should reduce or eliminate any of the protections for research subjects set forth in this [World Medical Association] Declaration" (World Medical Association, 2008)

5.1 European Data Protection Legislation

Currently the central legislative instrument of European data protection law is the "Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data", the so-called "Data Protection Directive" (95/46/EC). The objectives of the Directive are to ensure the protection of "the fundamental rights and freedoms of natural persons, and in particular their right to privacy with respect to the processing of personal data" (95/46/EC, Article 1-1) by the EU Member States and to achieve harmonisation of data privacy laws throughout the EU by developing and specifying data protection principles. The Directive is designed to protect the privacy of individuals and ensure the protection of personal data of citizens of the EU, especially with regard to processing, usage and transfer of such data. On these grounds, the Directive had to be transposed into national law by all EU Member States by the end of 1998. However, while the Directive includes a minimum set of provisions to be implemented by the Member States, the Member States are free to 'increase' the level of data protection for their country. The scope of the Directive encompasses

> "*the processing of personal data* wholly or partly by automatic means, and […] the processing otherwise than by automatic means of personal data which form part of a filing system or are intended to form part of a filing system." (95/46/EC, Article 3, Paragraph 1).

In the context of the Directive, *personal data* is broadly defined and refers to

> "any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly, in particular by reference to an identification number or to one or more factors specific to his physical, physiological, mental, economic, cultural or social identity" (95/46/EC, Article 2a).

Processing of personal data according to the Directive includes

> "any operation or set of operations which is performed upon personal data, whether or not by automatic means, such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction" (95/46/EC, Article 2b).

Therefore the provisions of the EU Data Protection Directive apply to SSH research data when data collection and processing include personal identifying information. However, since the Data Protection Directive (by definition) is not a self-executing legal instrument,

but leaves the choice of form and methods to the national authorities[8] (which almost inevitably includes differing levels of implementation), data collectors and depositors in the SSH domain have to consider the concrete legal framework within which their empirical research operates throughout Europe (cf. CESSDA); i.e.: European SSH data collection and processing has to comply with national and regional data protection law.

General provisions of Directive 95/46/EC, which in their national transposition are of relevance for SSH research when collecting and processing personal data, are for example:

Article 6, Paragraphs 1b and 1e, according to which

> "Member States shall provide that personal data must be […]
>
> (b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes. Further processing of data for historical, statistical or scientific purposes shall not be considered as incompatible provided that Member States provide appropriate safeguards; […]
>
> (e) kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed. Member States shall lay down appropriate safeguards for personal data stored for longer periods for historical, statistical or scientific use." (95/46/EC)

Article 7a and 7f, in accordance with which

> "Member States shall provide that personal data may be processed only if […] the data subject has unambiguously given his consent; or […] processing is necessary for the performance of a task carried out in the public interest […]." (95/46/EC)

Article 10, which states that

> "Member States shall provide that the controller or his representative must provide a data subject from whom data relating to himself are collected with at least the following information, except where he already has it:
>
> (a) the identity of the controller and of his representative, if any;
> (b) the purposes of the processing for which the data are intended;
> (c) any further information such as
> - the recipients or categories of recipients of the data,
> - whether replies to the questions are obligatory or voluntary, as well as the possible consequences of failure to reply,
> - the existence of the right of access to and the right to rectify the data concerning him

[8] Cf. Treaty on the Functioning of the European Union, Article 288 (ex Article 249 TEC).

in so far as such further information is necessary, having regard to the specific circumstances in which the data are collected, to guarantee fair processing in respect of the data subject." (95/46/EC)

However, it is noted that even after the implementation of the provisions of the Directive 95/46/EC there are still many differences between EU countries regarding national data protection laws/regulations, and sometimes even differences on the regional level can be observed in that regard (such as, for instance, between the German "Bundesländer"[9]). Since the provisions of the Directive have been implemented in different ways in the Member States, as a result, differences in the level of data protection, both on paper and in practice, exist.[10] Regarding this, it must be said that the Directive has failed to achieve proper harmonisation due to the different implementations of its provisions in the EU Member States.

In order to ensure a harmonisation of the legal framework and the legal practice within the EU, a new Proposal for a "General Data Protection Regulation" ("Regulation of the European Parliament and of the Council on the protection of individuals with regard to the processing of personal data and on the free movement of such data") has been released in January 2012. In contrast to the Directive, a Regulation is binding in its entirety and directly applicable in all Member States.[11] The General Data Protection Regulation proposed can reduce the existing fragmentation of data protection laws across Europe. For transnational surveys such as SHARE and ESS, for example, an important aspect in this respect is the standardisation of consent procedures across the Member States of the European Union (e.g. concerning the linkage of survey data with administrative data, where, for instance, currently in Denmark no consent has to be asked from the participants while in Germany asking for written consent is obligatory). It is hoped that the Regulation will be adopted in 2014 and that it takes effect after a transition period of two years in 2016.

Currently the provisions of the Regulation are subject of controversial discussion. Several amendments have been proposed.[12] Due to these on-going negotiations uncertainty persists regarding the implications of the Regulation with regard to research data generation and management in the SSH domain and the extent to which the Regulation will affect the work of the existing SSH research infrastructures. This as well as possible negative

[9] In Germany in addition to the Federal Data Protection Act ("Bundesdatenschutzgesetz"), each German state ("Bundesland") has its own data protection law.

[10] The website of the PRIVIREAL project provides information on how each Member State has implemented the Directive, in particular in the area of medical research. It provides the pertinent data protection laws and regulations for each country, as well as commentaries and other background information: http://www.privireal.org/content/dp/countries.php.

[11] Cf. Treaty on the Functioning of the European Union, Article 288 (ex Article 249 TEC).

[12] Cf. "Draft report on the proposal for a regulation of the European Parliament and of the Council on the protection of individual[s] with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)" of the Committee on Civil Liberties, Justice and Home Affairs (16.01.2013) and the subsequent amendments to the Proposal of a General Data Protection Regulation.

consequences for SSH research resulting from these debates are further discussed in Section 8.3.

Besides the Data Protection Directive and the anticipated General Data Protection Regulation, the "Directive on privacy and electronic communication" (2002/58/EC) is another important legal instrument for the protection of personal data in the EU. Directive 2002/58/EC complements Directive 95/46/EC with respect to the processing of personal data and the protection of privacy in the electronic communications sector. It has been amended by the so-called new "e-Privacy Directive" (2009/136/EC) (cf. ESOMAR, 2012), which extended the provisions of Directive 2002/58/EC in particular by introducing an obligation to obtain consent for the placement of cookies on websites.

The EU online privacy legislation and associated national laws also have consequences for the collection and the processing of data in the SSH domain; especially with regard to online research. When administering web surveys, for example, SSH researchers need to ensure that they are fully complying with the Directives and the associated national law in all EU countries in which they are carrying out their survey. Even though regulations under these laws initially may be intended to limit online behavioural tracking, they may directly affect the way in which data can be collected or even if/which data (e.g. specific kinds of paradata) may be recorded, depending on the rules adopted by each country.

5.2 Intellectual Property Rights: Copyright and Data Ownership

Besides data protection law, the second area of law that is of particular relevance for the conduct of SSH research is the area of intellectual property rights (IPR), in particular *copyright* and *database rights*.[13] The RESPECT Code of Practice for Socio-Economic Research states that the "material used in socio- economic research is predominantly protected by intellectual property rights such as copyright, database and software protection" (RESPECT, 2004). It recommends that "wherever practicable, intellectual property rights should be explicitly addressed in contracts covering the conduct of socio-economic research, whether these are funding contracts, partnership agreements or employment contracts" (ibid.).

According to the European IPR Helpdesk, intellectual property (IP)

> "protects the intellectual creation, not physical objects or facts. Therefore IP protects the information within the data in its original expression through copyright and in the investments for its collection through database rights, the inventions through patents, etc. Facts and ideas can be protected by keeping

[13] Key texts drawn upon in this section are Bainbridge (2007) Intellectual Property Law and the RESPECT Code of Practice for Socio-Economic Research; besides, advice from the European IPR Helpdesk (pers. com. July 2013) is cited.

them in secrecy (confidential business information and trade secrets)." (EC IPR Helpdesk, pers. com. July 2013)

Intellectual property rights issues are therefore linked to data harvesting, data access arrangements, data re-use and data linkage practices. These in turn are covered by data protection legislation (including the proposed Data Protection Regulation) and the commitments provided to respondents as part of securing informed consent.

For individual researchers this means that IPR must be dealt with in various stages throughout the research process. The RESPECT Code identifies the following as central to considerations of Intellectual Property: securing necessary permissions for the use of copyrighted material, correct attribution of authorship, acknowledgement of sources, correctness of references and the avoidance of plagiarism.[14]

Copyright and database right are the legal aspects of IPR that are of especial relevance to the SSH RIs, and are governed by a range of legal regimes, including the Berne Convention[15], EU and national legislation.

The SSH ESFRI RIs are all facilities which involve knowledge-based resources, i.e. "collections, archives or structures for scientific information" (Council Regulation, EC 723/2009: 4). The development of these knowledge-based resources from different sources means that the legal entity in charge of the RI assumes specific responsibilities and rights. Questions relating to data ownership are determined by reference to the legal arrangements covering data collection and data depositing and concerning issues of data protection and privacy.

For copyright protection to apply, databases must have originality in the selection or arrangement of the contents. However, if the copyright of the database contents belongs to a third party, this right will remain with them. For database right to apply there must have been a substantial investment in obtaining, verifying or presenting its contents. Often a database will satisfy both requirements.

Parry and Mauthner (2004) note that "the different disciplinary uses of data may have fundamental implications for issues of ownership, anonymity and consent." In respect of qualitative data they suggest that "because the construction of qualitative data is a joint endeavour between respondent and researcher, both parties should retain authorship/ownership rights over the data [with] very practical legal and ethical implications […] involving copyright and ownership." (ibid.) As a practical solution, they suggest that "copyright owners can waive their rights by assigning the copyright elsewhere or by

[14] The RESPECT Code also devotes attention to authorship. This is excluded in the current discussion – it is noted, however, that discipline based associations, e.g. the British Sociological Association, have guidance on authorship as well.

[15] Cf. the Berne Convention for the Protection of Literary and Artistic Works, an international agreement governing copyright.

licensing others to use their work while retaining ownership themselves" (ibid); this may allow a researcher to use the data in particularly ways in the absence of a written agreement as it may be demonstrated that the purposes for which the data would be used were "reasonably expected" (ibid.: 142). This example illustrates what is meant by the following RESPECT Code recommendation:

> "if a planned activity is not clearly covered by statutory permissions […] identify the rightsholder and conclude authorising contracts (transfer/assignment of rights/license agreements). Ascertain that the permission covers explicitly all relevant aspects – among them the description of type, extent, duration, environment (such as online) of the intended use […]." (RESPECT, 2004)

The RESPECT Code recommends that researchers should start from a position "assum[ing] that any material created or used in socio-economic research might be intellectual property and consider protection before using it" (ibid.).

The challenges associated with IPR of research data include the need for clarity as to the legal status of the legal entities involved; for example, when centre A transfers data to another, centre B, the recipient, will own the physical data, but that does not mean that centre B will own the copyright or other intellectual property rights. In the absence of an agreement, the creator of data will generally own the IP of the data even if he or she has transferred the data to someone else.

Therefore, for the SSH RIs there is a clear need to ensure that the relationship between the depositors (i.e. data collectors) and the knowledge-based resources such as archives or data centres is regulated by agreements which include regulation of the data use and intellectual property. This can for instance be done by using Depositor's Licence Agreements and End Users Licence Agreements.[16] The EC IPR Helpdesk suggests that "best practice is to regulate this matter in the contract and not to merely rely on the law […]." (EC IPR Helpdesk, pers. com. July 2013). Moreover, since copyright grants certain exclusive rights to the owner, such as producing copies and reproductions and creating derivative works, it is essential to use agreements that permit sharing and re-use of research data without copyright infringement.

Concerning the question on service agreements (i.e. subcontracts between commissioning bodies and data collectors, e.g. a research council commissioning a fieldwork agency to carry out a survey), the IPR Helpdesk suggests that it is essential to understand that the general principle of ownership usually applies in these agreements, that is, the creator is the owner of the IP in the commissioned work (although there are some exceptions to this in different countries). Therefore, it should not be assumed that in a subcontract the IP will belong to the commissioner. For this reason, it can be regarded as extremely important to have an agreement in place transferring all the results of the service, including all

[16] For an application of this approach, see Losnegaard et al. (2013).

intellectual property rights to the commissioner (EC IPR Helpdesk, pers. com. July 2013). According to the IPR Helpdesk,

> "[t]hus, it is advisable to agree on the ownership of data and intellectual property rights before starting research projects. This means that if there is any copyright in the data, metadata and paradata it will rest with the author, except if there is any legal exception or contract establishing otherwise." (ibid.)

Another aspect that is of importance for the SSH ESFRI RIs, is to make available clear statements about the terms of use of the data resources in each facility in order that the intellectual property rights of the partner institutions involved in a project are clearly defined. A Guide to Intellectual Property Rules for FP7 projects is available from CORDIS. This usefully provides overviews of IPR concepts such as 'background' (what participants bring to a project) and 'foreground' (what is generated as a result of the project activities) (see: ftp://ftp.cordis.europa.eu/pub/fp7/docs/ipr_en.pdf).[17]

Furthermore, as previously stated, copyright legislation may present challenges to the sharing and re-use of research data. While licensing and other agreements between data centres and data depositors as described above may provide useful solutions, the growing movement for open access to research data, coupled with initiatives to further the use of research exceptions to copyright legislation may be more fruitful as long-term solutions. This needs to be followed up by awareness-raising among the research communities.

In fact, the report of the European Commission on the consultation on the development of a framework for the European Research Area (ERA) reported, in the section dealing with 'overcoming the barriers to enhanced knowledge circulation through open access', that responding to a question on barriers to enhanced knowledge circulation through open access to publications and/or data in the ERA, respondents noted that the most important barrier is "the insufficient awareness of researchers on open access to data" (European Commission, 2012a: 39). Moreover, the Report observed a number of other obstacles, such as insufficient Member State policies on open access to data, insufficient awareness of researchers on open access to publications, low coordination of Member State policies, insufficient Member State policies on open access to publications and insufficient pan-European e-infrastructure for depositing scientific publications and data and EU copyright rules. (cf. ibid.) In general, the respondents identified the need to increase IPR awareness among researchers.

[17] It is noted that for most researchers, their terms of employment will include IP arrangements – in the UK, for example, the research data is the property of the employer, but this may be modified by, for example, provisions in a grant agreement between an institution and a funding body.

6 Identifying New Ethical and Legal Issues in the SSH

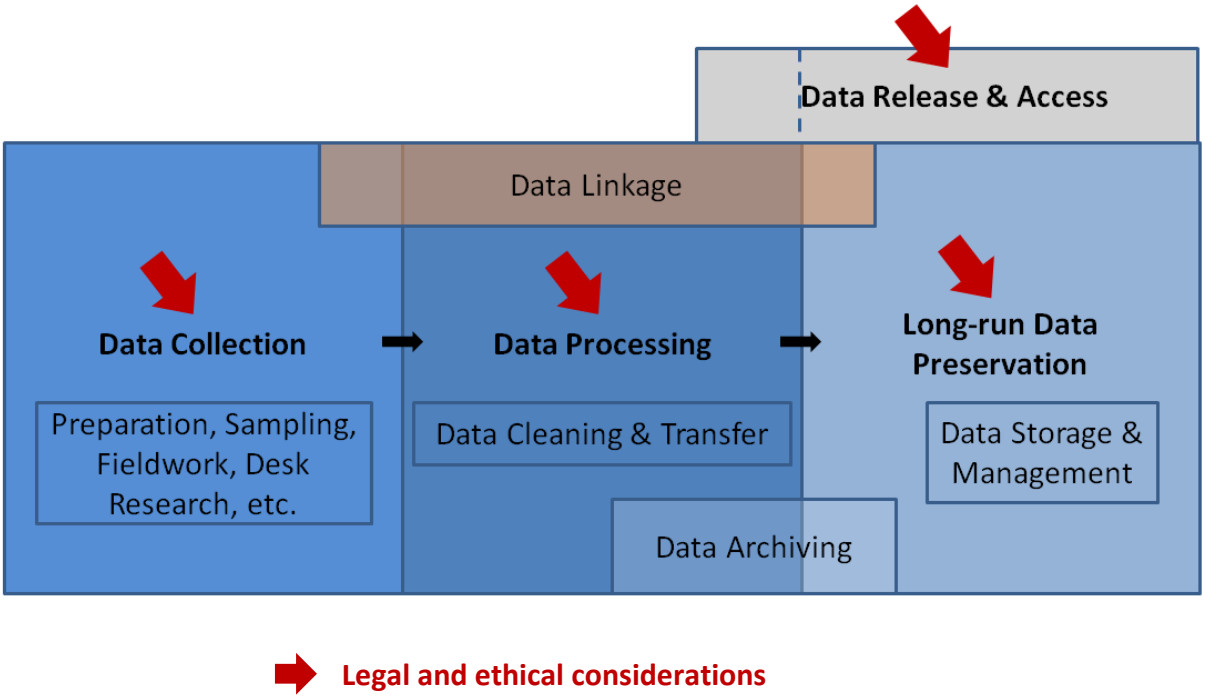
In accordance with Denscombe (2002) "consent, authorisation, integrity and data security" can be identified as central to the recognition of the rights and interests of participants/informants. These issues are generally accepted as key ethics issues and collectively they constitute the framework within which all data collection and data management issues may be considered. The legal framework relates to the data protection and data transfer regulations in place (cf. Section 5).

Kimmel (1988: 36) suggests that "ethical problems can be categorised and contrasted according to the level of the research process that they most directly affect", and while his typology moves from the individual to the society, this conceptualisation is of interest in focusing attention on the *entire research process* – it is suggested that there are effects in terms of focus – whereas earlier pronouncements that the stage of data collection should constitute the principal site to which scrutiny for ethics should be directed (cf. Bulmer and Warwick, 1993), the position currently advanced is that that this focus is now extended from the research data collection, to the data curation and data linkage capabilities and to data dissemination practices (cf. Section 3.3).

6.1 Different Phases of Data Generation and Management

Accordingly, when identifying and examining the legal and ethical challenges confronting researchers in the SSH domains, not only the different types of data, such as biomarkers, paradata, social media data or linked administrative data, to which these challenges are related have to be taken into account, but also – and even more importantly – the different stages of the research process, including data generation and data management have to be considered. The 'new' legal and ethical challenges experienced by researchers when dealing with these types of data, always take place within concrete operations performed as part of the SSH data collection, curation and dissemination process.

Ethical and legal issues, such as the issue of informing respondents about data collection purposes and data protection measures and obtaining their consent or the issue of anonymising data, occur at specific stages of the research process. Some issues occur only once during the research process, such as obtaining the respondents' consent, while others, such as implementing appropriate data protection measures, occur at different stages in varying forms – e.g., once as part of the information that has to be provided to the respondents prior to data collection, and once when actually being implemented during the processing of the data. The following figure gives an overview of the different stages of the research process in which legal and ethical aspects have to be considered:



6.2 Overview of Key Issues

The table below shows key ethical and legal issues in relation to the different stages of the research process (e.g. obtaining informed consent in the phase of data collection). Furthermore, it indicates which of the SSH ESFI research infrastructures (CESSDA, CLARIN, DARIAH, ESS, SHARE) are affected by these issues in the different phases of the research process and points out some specific "use cases". The "use cases" illustrate examples of types of data that have gained increasing importance in SSH research and associated ethical and legal challenges that have arisen. Specific ethical issues and legal requirements related to these "use cases" and the new challenges arising from them in cross-country SSH research will be further discussed and illustrated in Section 7 of this report.

Stage of the Research Process	Key Ethical (E) and Legal (L) Issues	Relevant to …	Special "Use Cases" / Examples
Data Collection (incl. Preparation Phase) (Gathering SSH data, sampling, fieldwork/desk research/web-	- Informed consent (E&L) - Ethics review (E) - Sampling (E) - Over-research (E) - Assuring anonymity and confidentiality (L&E)	CLARIN, ESS, SHARE	- Collection of dried blood spots - Web-based data collection (e.g. online surveys, social

based research)	- Survey agency contracts (incl. data ownership) (L)		media data) - Collection of spontaneous speech data
Data Processing (a): Data cleaning (Preparation of data for analyses, data cleaning, post coding, etc.)	- Anonymising/Pseudo-nymising data (L&E) - Ensuring that linkages (leading to identification of data subjects) cannot occur with ancillary data sets (L&E)	CESSDA, CLARIN, ESS, SHARE	- Data linkage with administrative records - Other linked/ heterogeneous data sets (e.g. including geo-referenced data)
Data Processing (b):Transfer of data from point A to archive (Data submission and transfer across borders – if applicable)	- Data protection/ confidentiality (L&E) - Encryption arrangements (L) - Data transfer licences (L) - Licence and acquisition agreements between data producers and archive/repository (L)	CESSDA, CLARIN, DARIAH, ESS, SHARE	- Transnational research (e.g. data linkage, transfer of biological samples) - Archiving of secondary data
Data Access and Long-term Preservation (Data storage and access arrangements)	- Data sharing licences & re-use purposes, data access regulations (L&E) - IPR/Copyright (L) - Security of data (L&E)	CESSDA, CLARIN, DARIAH, ESS, SHARE	- Data access and re-use of different types of paradata - Language data derived from copyrighted material - Text data mining of websites
'Dissemination'	- Open access arrangements (L&E)	Entire SSH,	- Online availability of

Phase (Phase beyond, presentation/ publication of data)	- IPR status of the derived resource/final resource (licensed for end use with e.g. Creative Commons or other licences) (L&E)	Research communities	articles, etc. based on previously gathered data

7 New Legal and Ethical Challenges in the SSH

The extent to which there *are* 'new' ethical and legal issues requires a nuanced approach. On the one hand, there are new ways of data collection (via blogs et cetera) and new ways of linking survey data to administrative data which raise issues about data security and anonymity. In this connection, the new technologies allowing unprecedented levels of data collection, data collation and data dissemination, amplify, rather than necessarily create, ethical challenges. On the other hand, it is possible to identify specific legal challenges which are novel. These raise issues of ownership, of legal right and legal liability and responsibility which are directly related to the Internet and new technologies (such as those associated with 'Big Data').

This section provides an overview of some central ethical and legal challenges related to certain types of data and/or new ways of data collection, linking and sharing in a cross-country usage scenario arising at the different stages of SSH research processes. The focus lies on issues which concern the SSH ESFRI research infrastructures currently in their day-to-day operations of SSH data collection, curation and dissemination. For each of the three main stages of the research process (cf. Section 6.1), first a general ethics issue and/or legal requirement entailing certain special challenges is discussed and then specific challenges that are of relevance for (one or more of) the SSH ESFRI RIs are examined.

7.1 Data Collection and Data Harvesting

7.1.1 Informed Consent (SSH)

For SSH researchers, and especially for survey researchers, two key ethical principles are to prevent respondents from harm and to assure the autonomy of the 'human subjects' of research (cf. Singer, 2008: 85; Couper and Singer, 2013: 57).[18] In practice, this involves ensuring the confidentiality of the data collected or harvested from human subjects (e.g.

[18] Besides the key issues of informed consent and confidentiality protection, the principle of 'justice' is advanced for the conduct of research involving human subjects. However, this principle, which aims at a fair balance between the subjects who bear the burden of research and those who benefit from it, is more important to biomedical research (cf. Couper and Singer, 2013: 57; Singer, 2008: 80).

respondents in survey research such as SHARE and ESS) and obtaining informed consent of them[19]. The second issue, obtaining informed consent of respondents, participants, informants, etc., mainly relates to data collection and data usage.

"[O]btaining respondents' informed consent […] has nothing to do with protecting subjects from harm, and everything to do with assuring that they are treated as autonomous individuals with the right to make informed, voluntary decisions about participation" (Couper and Singer, 2013: 57). To ensure this, usually consent – whenever it has to be obtained, whether in a written or a verbal form – has to be obtained prior to data collection. According to Singer,

> "[i]nformed consent requires (a) providing enough information about potential benefits and risks of harm to permit subjects to make informed participation decisions; (b) assuring that the information is understood: and (c) creating an environment that is free from undue influence and coercion. In addition, (d) research organisations ordinarily need some evidence that subjects have, in fact, been adequately informed and have agreed to participate" (Singer, 2008: 85).

"Included among the elements of informed consent are a description of the purpose of the research, the benefits and potential harm of participation, confidentiality protections provided, and the voluntary nature of participation" (ibid.: 84). In this respect, the question of how much and how detailed the information given to the subjects should be in order to ensure an adequate level of information on the part of the subjects poses a major challenge for SSH researchers (cf. ibid.: 86). Moreover, obtaining consent itself may be challenging in the context of certain research scenarios, such as the collection of spontaneous speech from recorded conversations, which does not only involve the primary human subjects, but also third parties (including those that may be discussed or mentioned), which have not been able to give their consent to the data collectors.

Even though obtaining informed consent is nothing new in SSH research, special issues arise, for example, when collecting and using paradata in the context of survey research (see Section 7.3.2 of this report), when harvesting data from the Internet or when spontaneous speech or text is recorded for research purposes. Especially with regard to the task of obtaining informed consent, it becomes obvious that new technologies allowing unprecedented levels of data collection, data collation and data dissemination, amplify ethical challenges and give rise to specific issues which are not covered in existing ethics codes.

For example, with regard to the issues of how survey participants should be informed about the collection and use of paradata and of how much information on this should be provided

[19] It is noted that "obtaining respondents' informed consent […] has nothing to do with protecting subjects from harm, and everything to do with assuring that they are treated as autonomous individuals with the right to make informed, voluntary decisions about participation" (Couper and Singer, 2013: 57).

to them, existing codes of ethics are not very clear (cf. Couper and Singer 2013). Moreover, from a legal perspective, in many cases of paradata collection it is not clear under which conditions specific kinds of paradata can be collected and how they may be used and released. Taking into consideration that the quality of surveys (which itself can also be considered as an ethical issue – cf. Singer, 2008: 96) depends on the response rates they achieve, informing respondents in an appropriate way, while at the same time avoiding a decrease of participation rates turns out to be a challenging task for survey researchers.

Furthermore, in some cases, such as the collection of biological samples (e.g. the collection of dried blood spots) or the linkage of data sources (e.g. survey data with administrative data), additional informed consent may be required.[20] For instance, when linking survey data with administrative record data is intended, a person who may already have consented to the participation in the survey is confronted with a request for additional data which may make consent necessary. Additionally, in relation to longitudinal surveys such as SHARE, which "often wish to carry out repeated linkages over time, there are ethical issues about the longevity of consent that is obtained" (Calderwood and Lessof, 2009: 68). However, even in these cases, Fulton states, "there do not appear to be any widely accepted 'best practices' for soliciting permission to access respondent records" (2012: 16).

Moreover, the current European data protection legislation is fragmented in that regard – concerning consent for data linkage no uniform procedures exist: while, for instance, in Denmark currently no consent has to be obtained when linking survey data and administrative record data, in Germany written informed consent is obligatory.[21] This fragmentation is not only causing uncertainty in the context of cross-country research but also forces SSH researchers to develop consent procedures on a case-by-case respectively a country-by-country basis.

7.1.2 Dried Blood Spot Collection in Population-based Surveys (SSc)

In the social sciences, an example of collecting innovative variables in population-based surveys associated with 'new' ethical and legal challenges is the collection of biomarkers. Biomarkers are objective measures of biological functions and therefore go beyond the subjective perceptions self-reported information (usually collected in the course of surveys) is based on. In many clinical and epidemiological studies, biomarkers have been collected

[20] Granted that the respondents have to agree respectively have agreed to participate in the survey.

[21] While according to Article 11, Paragraph 1 of Data Protection Directive (95/46/EC) on "[i]nformation where the data have not been obtained from the data subject […], Member States shall provide that the controller or his representative must at the time of undertaking the recording of personal […] provide the data subject with at least the following information […]" – According to Paragraph 2, "Paragraph 1 shall not apply where, in particular for processing for statistical purposes or for the purposes of historical or scientific research, the provision of such information proves impossible or would involve a disproportionate effort or if recording or disclosure is expressly laid down by law. In these cases Member States shall provide appropriate safeguards." Considering that the Directive includes a minimum set of provisions to be implemented by the Member States, allowing the Member States to 'increase' the level of data protection for their country, these differences can be explained.

through venipuncture and whole blood analyses. Since such a methodology is extremely difficult to apply and expensive when used in field surveys such as SHARE, the method of dried blood spots (DBS) has been introduced to the research field as an alternative. DBS sampling means that several drops of blood are taken via a finger-prick (using a small, sterile lancet – just as it is done daily by millions of diabetic people) and collected on filter paper, which is then dried and shipped by ordinary postal mail to a laboratory, where it is stored in freezers for a longer period before thawing and subsequent analyses.

Regarding the inclusion of DBS in population-based surveys there are several ethical and legal issues that have to be taken into account during (respectively prior to) the collection phase. In particular two issues are relevant: Firstly, additional informed consent is required; and, secondly, ethics committee approval has to be obtained. In particular, since the collection of DBS is considered to be an invasive method – even if only minimally invasive – it gives rise to various concerns from all parties involved in the DBS collection (researchers, survey agencies, interviewers, participants) that not only have to be addressed, but also make ethics committee approval necessary.[22]

While the procedure of obtaining informed consent in this case[23] is relatively easy since the legal requirements in the EU Member States are consistently demanding informed consent in written form, and the information that has to be provided to the respondents, in principle, is not too difficult to identify, obtaining ethics committee approval from all participating countries in a transnational survey such as SHARE, however, is a real challenge.

Identifying the responsible ethics committee/s in each country and applying to them in accordance with the respective national or regional policies and procedures is in itself a very challenging task since they are highly fragmented across European countries. However, not only does the organisation of the national ethics committee systems differ a lot between the different EU Member States, but also the requirements of these ethics committees vary a lot. Therefore, meticulous preparations of the applications for ethics committee approvals have to be carried out.

For this reason, as part of WP6 of DASISH a transnational systematic inquiry on national legal requirements and ethics committee approval procedures in the EU with regard to the collection of biomarkers (derived from DBS samples) has been carried out making use of the SHARE Research-Network. Part of this inquiry was a set of questions, including legal and ethical issues to be addressed in the course of the ethics reviews; these are presented in Annex 10.4. In accordance with the first findings of this inquiry, the following legal and

[22] Besides the issues occurring during the data collection phase, it is noted that – if variables that are derived from biological samples are included in social science data sets – especially processing and transferring of the samples as well as the data derived from these appears to be challenging (e.g. separation of blood samples and personal data has to be ensured during all phases of processing and storage). Furthermore, the collection of biological samples, such as dried blood spots for biomarker analyses in SHARE, changes the nature of the database: the database becomes a biobank (requiring specific biobank regulations, etc.).

[23] In contrast to data linkage, where no uniform consent procedures exist (see Section 7.1.1).

ethical issues (occurring during different phases of data generation and management) that require special attention when applying for ethics committee approval for the collection of DBS in the course of a transnational population-based survey can be summarised:

- Obtaining informed written consent of the respondents; i.e.:
  - Ensuring that participation is absolutely voluntarily (including providing options for withdrawal) and
  - Providing adequate information to respondents (here, possible specific national or even local legal requirements should be taken into account)
- Defining the research context and specifying the rationale, methods and objectives of the DBS collection (this is very important with regard to the classification of the study made by ethics committees, e.g. as a clinical-epidemiological or multi-centre study, which might substantially influence the final requirements of the ethics committee/s)
- Collection of DBS samples by trained interviewers (in some countries – e.g. in Austria and the Czech Republic – only medical personnel may collect the DBS samples)
- Addressing the various concerns from all parties involved in the DBS collection in appropriate ways and providing training (if applicable) and all relevant information to them; making additional insurance arrangements, if necessary
- Providing feedback to the participants about the results of the analyses conducted (here the opinions and requirements of different ethics committees may vary substantially: some ethics committees demand that participants have to be informed about the results of the DBS analyses – sometimes only via a general practitioner, other committees demand exactly the opposite, i.e. not to inform the participants)
- Ensure data privacy of the respondents with regard to the shipping and storage of the biological samples and the related consent forms (e.g. in Switzerland shipping of consent forms across borders is not allowed)
- Ensuring data protection when setting up a biobank, including the biomarker database and the linking of the analyses results to the survey data set

7.1.3 Collection and Harvesting of Language Data in the Humanities (Hum)

In the humanities domain, in recent years, the language sciences and language technologies have developed a wide range of modelling approaches based on naturally occurring language data. These data are obtained from an increasing variety of sources. Resulting models have a substantial range of applications in information and communication technologies, language teaching, etc.

For example, recordings of spoken language data are studied to discover patterns of language use and form a basis for research and development in language and speech processing. Language models have a wide range of applications, such as dictation interfaces, natural language interaction (e.g. Siri on the iPhone), interpreter services, etc. In some cases language recordings are transcribed to make them searchable and analysable. Since spoken

language data may include personal information, some of which may be of a sensitive nature, several legal and ethical challenges may arise during the collection of these data. These challenges may vary according to the method of gathering data; in general they pertain to the privacy and autonomy of a) the informants and b) third party individuals that may be discussed in the data.

Primary spoken language data may be derived from a range of sources, such as natural daily conversations, telephone interactions, interviews, reading aloud, radio and television programs, etc. While recordings of people being instructed to read a text aloud usually are unproblematic from an ethical/legal point of view, since the content is controlled, these recordings are only useful to study phonetic aspects. In contrast, recordings of spontaneous speech are in many respects more useful, but are also connected to legal and ethical challenges. On the one hand, they are often produced by having informers carry with them a recorder to record conversations as they go about their daily business.[24] Even though in these cases, written permission can easily be obtained from the informants, ensuring the autonomy and privacy of third parties involved in these daily conversations poses a problem, since they have not been able to give their consent to the data collectors.

On the other hand, spontaneous speech data may also be generated from contexts that were not intentionally (or not only) geared towards language research. Here, obtaining consent prior to the collection of data whilst upholding the scientific value of the collected data may be a challenging task, since asking for consent might have an influence on the research context and therefore substantially change the outcome of data collection or even make data collection and therefore scientific research impossible. An example for such a data collection is the Nottingham Health Communication Corpus, which contains recordings of questions at a teenage health advisory service. In this case, no explicit consent for data collection was obtained from the speakers, but the speakers' identities were concealed and the speech was not made available to researchers other than a small group which transcribed it and produced aggregated linguistic analyses.

On a general note, such data collections may not only be problematic since it might be possible to retrieve informants' identities from the data (particularly if data have been collected in small communities), but also due to the lack of explicit consent there is a risk that national laws or rules put forward by ethical committees might be violated.

Another field potentially presenting 'new' legal and ethical challenges to researchers in the humanities is the collection and harvesting of written language data on the Internet, which has emerged as an important source of data for the language sciences and in text data mining (TDM). In general, harvesting previously published language data from the Internet is

[24] Examples of such corpora are the British National Corpus and the COLT corpus (The Bergen Corpus of London Teenage Language, see Stenström et al., 2002).

currently a legal and ethical grey zone, but legal initiatives are underway in some countries[25] (see also Section 7.3.3).

Besides previously published text on the Internet, text messages and social media also present new and useful sources for language data as they represent instances of spontaneous writing and may offer insights into writing practices and language patterns in informal genres as well as insight into the creation of new vocabulary. This kind of data is potentially available in large quantities and may be harvested by web crawlers (social media) or through crowdsourcing or informants providing their messages (SMS and social media). They do however present ethical challenges with regards to privacy and autonomy (cf. Swatman, 2012).[26] Especially in this area, with regard to the issue of obtaining informed consent, including related factors such as age verification and documentation, the new technologies used for data collection and data collation, amplify ethical challenges and give rise to specific issues which are not covered in existing ethics codes. For instance, it is easy to fake one's age when opening a Facebook account so verifying that users are 18+ could be problematic. And moreover, even if issues may be resolved with the informants, there may still be issues related to third parties mentioned in the language data, which have not been able to give their consent to the data collectors.[27]

7.2 Data Processing (incl. Data Transfer) and Data Linkage

7.2.1 Anonymisation and Pseudonymisation (SSH)

Ensuring confidentiality of data collected in the course of SSH research involving human subjects is of crucial importance since "most serious risks of harm to which participants in social research are exposed are breaches of confidentiality, and the consequences that may follow from such breaches" (Singer, 2008: 90). While issues such as obtaining informed consent and obtaining ethics committee approval primarily relate to the data collection phase or even are part of the preparation process of data collection – even though, of course, participants/informants also have to be informed about data protection measures taken in order to ensure their privacy prior to data collection – ensuring the confidentiality of research data becomes crucial in relation to data processing and data release and measures to ensure data privacy have to be implemented as part of the data processing.

[25] In Norway, e.g. an encompassing Norwegian web corpus has been compiled with the explicit permission of the Government.

[26] According to Swatman (2012), generally, research on social media should take the following issues into account: Recruitment, privacy/anonymity/confidentiality, consent (incl. age verification and documentation), data sharing and data storage and terms of service/end-user licence agreements.

[27] E.g., see Adolphs et al. (2011) for a presentation of a SMS corpora, where a group of nine informants have made the SMS logs available for researchers – the article does not explicitly state whether all senders of the messages have actually given their consent to the messages being stored in the corpus.

Whenever sensitive or confidential information[28] is collected and processed – which quite often is the case in SSH research[29] – there is a potential risk that this information will be revealed to unauthorised others, which might lead to negative economic, social, psychological consequences – such as the loss of employment, the loss of reputation, stigmatisation and discrimination or even criminal penalties. Therefore it can be considered as crucial not to disclose the identities of the participants/informants in SSH research. Whenever personal data – i.e. "any information relating to an identified or identifiable natural person ('data subject'); an identifiable person is one who can be identified, directly or indirectly" (95/46/EC, Article 2a; cf. Section 5.1) – is processed in the course of research projects, all measures necessary to ensure data privacy have to be taken. This is not only an ethical issue, but has to be done in the compliance with European and national/regional data protection law.

Amongst data collectors and processors, it is widely accepted that "a person's identity can be disclosed from:

a) *direct identifiers* such as names, addresses, postcode information, telephone numbers[, ID numbers,] or pictures

b) *indirect identifiers* which, when linked with other publicly available information sources, could identify someone, e.g. information on workplace, occupation or exceptional values of characteristics like salary or age"
(UK Data Archive)

Research data sets therefore should be checked for both direct identifiers and indirect identifiers prior to the use or release of the data. Furthermore, besides ensuring that linkage cannot occur with ancillary data in order to prevent statistical disclosure, the entire data environment has to be considered. Particularly, if the linking of different data sets is intended, additional attention has to be paid to *relational data*, i.e. to variables in both data sets that might lead to the disclosure of identities when being connected.

If SSH data include direct identifiers, indirect identifiers or relational data that might lead to a disclosure of the identities of respondents, appropriate measures have to be taken in order to ensure the confidentiality of the data (unless the respondent explicitly has given consent to use or release these data). That is to say, all technical and organisational measures as laid down in the relevant national (or regional) legislation of the Member States of the European Union that implement the European Data Protection Directive (95/46/EC) have to be taken.

[28] 'Sensitive data', on the one hand, can be understood as being of a particularly risky nature with regard to possible negative outcomes when being revealed to unauthorised others, such as information on racial or ethnic origin, political opinions, the sexual life or religious beliefs of the data subject. 'Confidential data', on the other hand, can be understood as information, which is protected against unwarranted disclosure for issues pertaining to personal privacy or for proprietary considerations.

[29] Cf. CESSDA.

Anonymisation and pseudonymisation of personal data are central security measures to ensure data confidentiality and the safeguarding of sensitive data and confidential information (cf. UKAN). They can in principle be applied at every stage of data processing. According to the European Data Protection Supervisor (EDPS)

> "[a]nonymisation of personal data means changing a data set so that it becomes impossible for the controller or for anyone else to identify a person to whom the data relate either directly or indirectly. Anonymous data are not personal data and fall outside the scope of data protection legislation. Anonymisation requires not only deleting all directly identifying attributes (e.g. names, civil registry numbers, phone numbers, biometric data) from the data set, but usually also data which in combination reveal unique characteristics and any further modifications, to prevent re-identifiability." (EDPS, 2013: 2)

Pseudonymising data, on the other hand means disguising identities, whilst retaining the possibility to backtrack to the individual under predefined circumstances. This, for instance, is necessary in panel studies, such as SHARE, that need to re-contact the participants of previous waves. The main difference between anonymisation and pseudonymisation lies in the way of treating direct identifiers – whilst in the case of anonymisation direct identifiers are completely removed from a data set, when pseudonymising data sets at least some of these identifiers are 'merely' replaced by a pseudonym (e.g. a coded number) while at the same time a means that may be used to identify the person (e.g. a decoding key[30]) is being kept by the data providers.

> "Retraceably pseudonymised data may be considered as information on individuals which are indirectly identifiable. Indeed, using a pseudonym means that it is possible to backtrack to the individual, so that the individual's identity can be discovered, but then only under predefined circumstances. In that case, although data protection rules apply, the risks at stake for the individuals with regard to the processing of such indirectly identifiable information will most often be low, so that the application of these rules will justifiably be more flexible than if information on directly identifiable individuals were processed." (Article 29 Data Protection Working Party, 2007: 18)

While direct identifiers are removed from data sets usually, whenever anonymisation of a data set is intended, with regard to indirect identifiers there are also other techniques for handling risk disclosure that can be applied in order to ensure the confidentiality of data. A

[30] Of course, this key – i.e. the connection between the pseudonym and the identifying data – has to be effectively separated from the data involved, and the identification by unauthorised persons has to be prevented effectively (cf. EDPS, 2013: 3). "Key-coded data are a classical example of pseudonymisation. Information relates to individuals that are earmarked by a code, while the key making the correspondence between the code and the common identifiers of the individuals (like name, date of birth, address) is kept separately." (Article 29 Data Protection Working Party, 2007: 18)

number of commonly used options when dealing with variables, which might act as indirect identifiers in quantitative data sets, are presented on CESSDA's website:

1. "Removal – eliminating a variable that contains direct identifiers entirely from the dataset. Remove, for example, respondent's names and addresses, postcode and so on.
2. Aggregation or reduction of the precision of a variable - reducing precision of potentially revealing socio-demographic by reducing the details of some characteristics, such as the respondent's age and place of residence. Or, record the year of birth rather than the day, month and year.
3. Bracketing – combining the categories of a coded (categorical) variable into a broader code. If using standard hierarchical codes (such as occupational codes), this process can be automated.
4. Top-coding – restricting the upper and lower ranges of a continuous variable. Salary, for example, is often top-coded to avoid identification of those with particularly high salaries.
5. Collapsing and/or combining variables – merging the concepts embodied in two or more variables by creating a new summary variable. This involves generalising the meaning of a nominal string variable; for example, specific types of training or qualifications which might identify particular respondents.

Other techniques, which should be carefully considered before they are implemented as they might result in some loss of analytical power for the data collection, are:

6. Sampling – releasing a random sample of sufficient size to yield reasonable inferences, rather than providing all of the original data.
7. 'Swapping' – matching unique cases on the indirect identifier, then exchanging the values of key variables between the cases. This retains the covariate structure while retaining the analytic utility. […]
8. Disturbing – adding random variation or stochastic error to the variable. This retains the statistical properties between the variable and its covariates, while preventing someone from using the variable as a means for linking records."

(Extracted from CESSDA's website, accessed 27/06/2013)

It is noted that, while "[p]rojects based upon a quantitative methodology can usually deal with [anonymisation] in a straightforward manner[, ... r]esearchers using qualitative methods need to approach the problem in a much more considered and reflective way" (CESSDA). Instead of crudely removing or aggregating data pseudonyms, replacement terms or vaguer descriptions have to be used in order to maintain the usability of the data (cf. UK Data Archive).

However, while direct identifiers, which, for example, are often collected in the course of survey administration, usually can be removed from the quantitative data sets easily, since these do not constitute information that is needed in the context of methodological or scientific research (cf. UK Data Archive), removing or modifying indirect identifiers or

relational data that could lead to disclosure of identities, also poses a more challenging task for SSH researchers dealing with quantitative data sets. With regard to these, anonymising or pseudonymising data might result in a loss of data usability.[31] In these cases alternative measures to ensure data privacy, such as access and usage restrictions, should be considered (see Section 7.3.1).

7.2.2 Linking Administrative Records with Survey Data (SSc)

In the social sciences, micro-level administrative data are increasingly being linked to survey data in order to provide researchers with richer databases and to open "a wide range of research possibilities for content related research as well as methodological research" (Korbmacher and Czaplicki, 2013: 47). This development "has been greatly facilitated by computerisation of administrative records and by technological advances" (Calderwood and Lessof, 2009: 58). In general, two major motivations for linking with administrative data can be identified: First, data from administrative records are usually much more detailed in comparison with survey data and provide a high degree of accurateness (often over time), so that these data enhance survey data with very useful information; and secondly, if the information contained in an administrative data set overlaps with the information that is included in the survey data set, this overlap of information can be used to validate the survey data.

Linking survey data with administrative record data raises several legal and ethical challenges. As Gill states, "[t]he matching of data should include consideration of the following:

- The ethics of linking the datasets, depending upon the source and content of the datasets.
- Confidentiality of data about individuals and businesses in the linked result set.
- Physical safeguarding of confidentiality including security of computer systems and administrative systems.
- […] compliance with the relevant legislation." (Gill, 2001: 100)

With regard to what Gill calls 'the ethics of linking the datasets', particularly the questions of whether and how (additional) informed consent should be obtained from data subjects (i.e. the participants in the survey) are of relevance. These issues are raised prior to data collection and linking, whereas, during data processing, including data transfer and linkage, ensuring the confidentiality of the data collected is the major ethical and legal issue.

However, to what extent this issue needs ethical consideration and to what extent legal provisions apply depends on the way in which the data are linked.[32] One the one hand, data

[31] "Of course, anonymising data makes them less useful than accurate, fine-grained data." (UKAN)

[32] In turn, according to Korbmacher and Czaplicki, "the method of linking different data sources depends on legal and technical constraints of each dataset." (2013: 49)

can be linked directly, i.e. by matching data sources of exactly the same person (so-called record linkage); on the other hand, linking can be done by matching data sources of people who are similar in a statistical sense (statistical matching). Whilst in the latter case in principle completely anonymised data can be used[33], in the first case – which from a scientific perspective often is the favoured method, since in the latter method the linked data only refers to a 'statistical twin' – identification of the participants is required in order to ensure (with a high probability) that the correct individual data sets are matched.[34]

Regarding the first case, in some European countries an obligation exists to ask survey participants for written consent (e.g. Germany) or to ask them for verbal consent (e.g. Austria) prior to the extraction of their administrative records and the subsequent linkage with their survey data. In this connection, in particular, the current fragmentation of European data protection law makes it difficult for transnational surveys such as SHARE to develop standard procedures for the linkage of survey data with administrative record data and necessitates investigations and preparations on a country-by-country basis (cf. Section 7.1.1).

Furthermore, when directly linking survey data with administrative records, it is of particular importance that safeguarding of personal data and confidential information is ensured in compliance with European and national data protection laws. Here, in particular at the stage of data processing and transfer, survey researchers are facing similar challenges – not only due to the fragmentation of data protection law in the EU Member States but also because of the varying provisions with regard to transfer and access to administrative record data of the respective national institutions which are providing administrative data for scientific research purposes.

Some European countries, such as Sweden and Denmark, which have integrated statistics systems that provide comprehensive population statistics databases, have a long history of using administrative data as a research resource. Since "[i]n these countries, integration of statistical data sets is a normal part of the operations of the national statistics office […, t]hese countries usually have a strong framework of legislation and clear rules about protection of confidentiality of personal and business data, irrespective of whether or not the data has been integrated from different sources" (UNECE, 2009). On the other hand, for other countries, such as Germany or the United Kingdom, "the notion of integrating data to produce composite microdata from different sources for statistical and related research purposes" (UNECE, 2009) and, in particular, the integration of administrative data into

[33] In this case, usually no (additional) consent has to be obtained from data subjects.

[34] When linking data on an individual level, two linkage methods can be used: exact/deterministic matching or probabilistic matching (cf. Calderwood and Lessof, 2009: 58-61). While probabilistic matching is based on several matching variables (which are allocated with different weights) and allows for disagreement between them, deterministic matching "depends on a unique identifier in both datasets" (Korbmacher and Czaplicki, 2013: 49), e.g. the respondent's social security number. In both cases, however, common features have to be present in the source data sets, in order to be able to link microdata from different sources (cf. UNECE, 2009). These common features necessarily refer to the data subjects – i.e. they are identifiers.

survey design is still relatively new (cf. Calderwood and Lessof, 2009). Especially in these countries the task of linking survey data and administrative record data presents 'new' legal and ethical challenges to researchers.

Safeguarding of personal data and confidential information particularly includes ensuring that the identity of respondents is

a) neither revealed in the course of the linking process,
b) nor as a consequence of the information included in the linked data set.

During the process of linking the data sets, data have to be exchanged between the institution holding the administrative records and the research institute conducting the survey. This imposes special challenges to survey researchers, which have to ensure that all measures necessary to safeguard data privacy are taken at every stage of the linkage process. When the data sets may only be linked in the institution providing the administrative record data, for example, a linkage procedure has to be developed that does not allow the institution keeping the records to match the survey data to individual administrative records that contain identifying information. In such a scenario, for instance, a solution may be to pseudonymise both data sets, but mark individual cases in both data sets with the same coded number, which allows them to be merged together later on for scientific research but effectively prevents de-anonymisation of respondents. The exact measures taken of course depend on the concrete linkage method applied and the technical constraints of each dataset and therefore have to be decided upon on a case-by-case basis – this can be an extremely challenging task, if data sets from several countries are to be linked to a survey data set as part of a trans-national research project, such as SHARE.

Furthermore, besides linking the data sets without disclosing the identities of participants, linked data sets themselves present a special legal and ethical issue in terms of the merged data they contain after being linked together. Since the details contained in a linked data set go beyond the details of each single data set, special attention has to be paid to relational data. Variables that might have been entirely unproblematic in terms of privacy and confidentiality in the initial data sets might lead to the disclosure of identities when being connected. Even though checking for variables of this kind can be an extremely difficult, depending on the size and the granularity of the linked data sets, this work has to be performed very carefully in order to effectively protect the privacy of individuals and to prevent respondents from potential harm. This especially has to be considered as well, when linking with administrative data is envisaged to be carried out repeatedly linkages over time and therefore the data set obtains a new long-term dimension (covering life courses eventually).

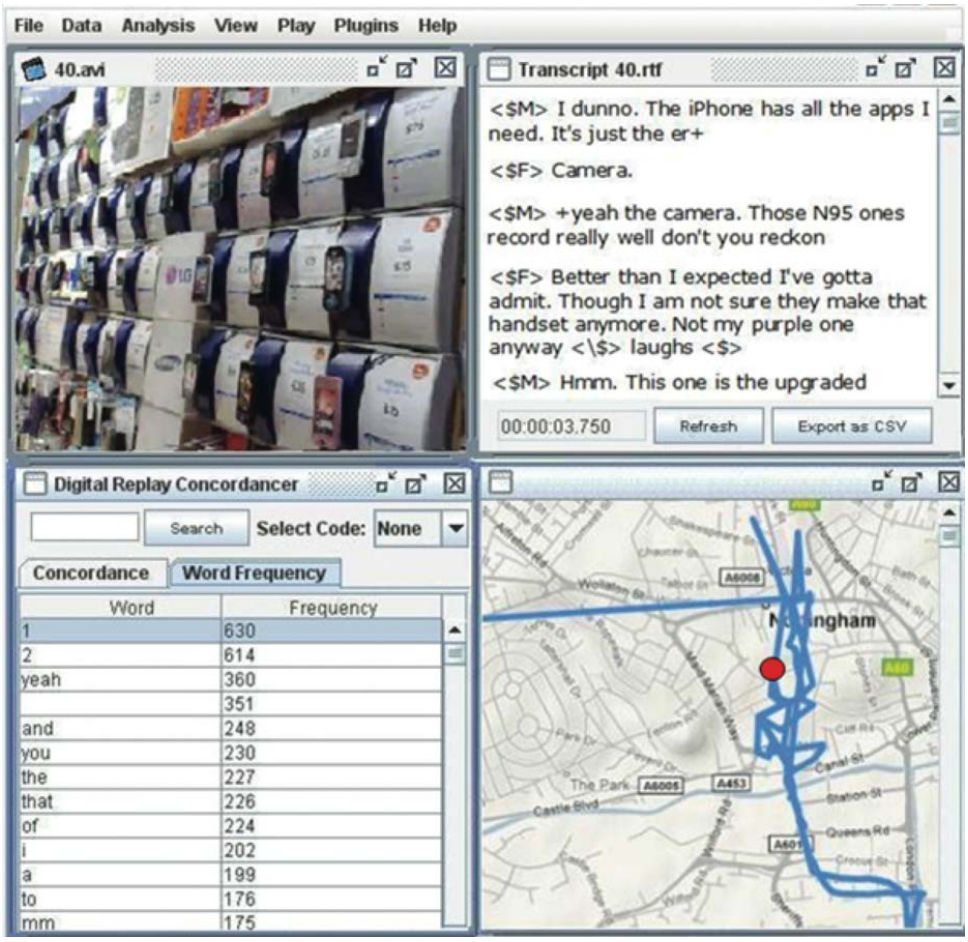
7.2.3 Privacy Issues Related to Language Data Collections (Hum)

Linguists are increasingly taking into account the speakers' environment when analysing language use. Moreover, technological advances have changed people's communication patterns as well as linguists' access to new technologies for documenting language use. Adolphs et al. (2011) present research that correlate language use with aspects of the speakers' environment collected from sensors that provide data on position, movement, time, etc. "to allow for the exploration and analysis of the patterned use of words, phrases, extra-linguistic and metadata information within and across devices and/or data type(s), time and/or location and participants/contributions" (ibid.: 310). In this connection, the linking of different types of data into the corpus design presents a number of legal and ethical challenges, which Adolphs et al. summarise in the following manner:

> "[ethical considerations] fall into the following broad categories:
>
> - Institutional: Guidelines prescribed by a particular institutional authority, University (imposed by a central Ethics Committee) and/or department.
> - Professional: Common guidelines used across a specific discipline, research paradigm and/or research funding council.
> - Personal: Personal and/or collaborator defined ethical standards which exist to maintain relationships and integrity in research.
>
> Moral and legal obligations faced at each of these levels can heavily influence processes undertaken during every stage of the corpus development, from the data collection phase through to its presentation and analysis. They may also vary across international dimensions which may have an impact on the ability to share or co-develop data and tools." (2011: 319)

Furthermore, technological advances make it possible to aggregate speech, camera, GPS and possibly other personal data, viewable through programs such as Digital Replay System[35] (see figure on the following page for an example of viewing of aggregated data). Linking language data to information about geographical location obviously adds further privacy issues to those already associated with spoken corpora. Amongst other issues – since "[r]ecent research suggests that also fine grained location data can be sufficient by itself to identify the individual it relates to" (EDPS, 2013: 2) – aggregated data sets including GPS data, fall within the scope of data privacy legislation and have to be handled in compliance with European and national/regional data protection law.

[35] "Digital Replay System (DRS) is a software tool being developed by the DReSS node of the UK ESRC-funded National Centre for e-Social Science. It is publically available under an open source license and is designed to support the organisation, synchronised replay, and analysis of complex multimodal corpora including audio, video, dialogue transcripts and system log files." (Brundell et al., 2008)



Viewing of aggregated data, including speech, camera and GPS (Source: Adolphs et al., 2011: 317).

Another development that shows how the use of new technologies in the humanities amplify ethical challenges and even give rise to new legal challenges relates to recent advances in the distribution of written corpora via digital infrastructures. This is illustrated[36] in the Slovene corpus "Nova beseda", containing 318 million words from newspapers, magazines, books etc. and available for web search, accessible without authentication. In July 2012, Slovenian authorities decided that all personal names in the corpus should be either anonymised or excluded from the results in the online search facility for data protection reasons. After negotiation, the corpus owners were allowed to provide searches for a name, but not for a combination of names (and/or surnames), although this is publicly available data. According to the Slovenian authorities, the corpus is a new structured collection of personal data, and cannot be compared with previous publication of the data for a different purpose. Moreover, the corpus provides much easier access to personal data than collecting it via newspaper articles stored in libraries etc. In addition to existing legal

[36] This case is based on information from Špela Vintar to ResearchGate.net and from Simon Krek to the Corpora list.

privacy requirements, in the future corpora including person names may represent a breach with the "right to be forgotten" that may be included in future EU data protection law.[37]

However, with regard to previously published written language data not only from a legal perspective but also from an ethical point of view privacy issues related to third parties mentioned in the language data can be identified. Sampson (2000), for example, argues that the interest of third parties should be protected through anonymisation, and even more so since third parties have not been able to give their consent to the data collectors. This may hold not only for spontaneous speech and text, but also for corpora of published texts. Sampson notes that in the case of the British National Corpus, names of well-known public figures or institutions have generally not been anonymised at all the references at all, and argues that such names too should be anonymised if the context is slandering: "Even American actresses, surely, are entitled to have their honour guarded by corpus linguists." (Sampson, 2000). As a result of this, in his CHRISTINE corpus Sampson has anonymised (replacing the name with the <name> entity) third party names in contexts "where it seems possible that the identification could cause embarrassment […]" (Sampson, 2000).

Sampson, however, represents one possible view; and as the example of the British National Corpus shows, no commonly accepted 'best practice' with regard to privacy issues related to third parties in connection with corpora of previously published texts exists.

7.3 Data Re-use and Data Access

7.3.1 Data Access and Usage Restrictions (SSH)

As has been shown in Section 7.2.1, anonymisation and pseudonymisation are central security measures to ensure data confidentiality. With regard to data re-use and data access completely anonymised (resp. pseudonymised) data sets have the advantage that they can be made accessible to the entire scientific community and even the entire public without restrictions on use or other conditions. Making use of this advantage, ESS data, for example, are publicly available almost without restrictions (cf. ESS conditions of use). The ESS ERIC allows free access to all data of the ESS for non-commercial use, scientific research, knowledge and policy making. In accordance with data protection regulations in the participating countries, correspondingly, only anonymous data are available to users.[38]

However, as pointed out in Section 7.2.1, anonymisation and pseudonymisation also have a significant disadvantage when being applied to indirect identifiers contained in scientific data sets: they make research data less accurate. As the UK Anonymisation Network states,

[37] Cf. The Telegraph, 25/01/2012: http://www.telegraph.co.uk/technology/news/9038589/Digital-right-to-be-forgotten-will-be-made-EU-law.html, accessed 30/06/2013.

[38] There are no privileged access rights by any person to the ESS data, except from what is necessary for its processing and preparation for public use.

"anonymisation […] always involves a trade-off between data utility and privacy-preservation" (UKAN). So-called 'public-use files', which by definition may only contain absolutely anonymised data, are limited with regard to their usability for scientific and/or methodological research, since some information has to be removed from them and some of the data contained has to be adjusted through data-masking procedures (cf. CESSDA). Even though, the degree to which usability of a data set is affected by anonymisation measures depends on the concrete characteristics of the data set, it is for this reason that alternative measures to ensure confidentiality, which at the same time maintain the usability of the data, should be considered when preparing SSH data sets that include sensitive and/or confidential information for use and re-use. Concerning this matter, CESSDA generally states:

> "Anonymisation is often the first approach considered by most researchers, but this should not be considered in isolation. Sensitive and confidential data may also be safeguarded effectively through access and usage restrictions employed in certain circumstances and if deposited in a formal archive." (CESSDA)

More sensitive and therefore less anonymous versions of the data, for example, may be effectively safeguarded when made available for scientific analyses to vetted users via 'on-site use' (i.e. analyses of data in separate secure workplaces for guest researchers) or 'remote data access' (i.e. indirect access to confidential microdata)[39]. Furthermore, usage restrictions, such as 'end user licences' can be used to safeguard sensitive and/or confidential data. On CESSDA's website the following alternative controls to protect confidentiality are listed:

- "The restriction of access by requiring users to sign up to legally binding conditions of use.
- Technological controls which prevent unauthorised users from accessing sensitive materials.
- Data enclave[s] or a secure data analysis [laboratories that allow] researchers access to the original data in a controlled setting.
- The creation of […] restricted-use data collections."

(Cited from CESSDA's website, accessed 27/06/2013)

Besides minimising the trade-off between privacy preservation and maintaining the usability of the data, data access and usage restriction have the advantage that they can also be used to enhance data protection. Taking into consideration that anonymising data, like any security measure, is not 'fool proof', in particular restricting access to sensitive or confidential data has a specific advantage in comparison to anonymising and pseudonymising data with regard to the prevention of statistical disclosure. Data access

[39] Remote Data Access (RDA) allows researchers to submit their own computer programs to research data centres (RDCs). At the RDCs, these will be run on the confidential microdata sets. Subsequently, after having been scrutinized for confidentiality, the results are returned to the researchers.

restrictions can effectively contribute to reducing risks of re-identification of individuals as a result of linking research data that has been stripped of personal data with other publicly available information sources (cf. Singer, 2007: 91, 94).

> "There is a huge difference between making data available to a small number of vetted individuals in a small lab and publishing the data as open data on the Internet. In the former case, opportunities for de-anonymisation are going to be very limited whereas in the latter any data anywhere in the world may be used to de-anonymise the data." (UKAN)

Furthermore, access to research data can be made subject to certain conditions of use, which also can be used to enhance data protection. Typical conditions respectively restrictions designed to augment anonymisation or confidentiality of research data include:

- "End user licence to respect confidentiality and not to disseminate any identifying information; a standard clause affecting all users of research data. Such a written undertaking does have contractual force in law. Furthermore, the good reputation of a secondary user depends upon abiding by these undertakings.
- Restricted access to certain kinds of highly sensitive data; for example, permission from the data creator might be required to access the materials." (CESSDA)

SHARE, for example, employs data access rules consisting of a combination of both data access and data usage restrictions (cf. SHARE data access rules). Firstly, applicants must have a scientific affiliation and have to sign a statement confirming that under no circumstances the data will be used for other than purely scientific purposes. And secondly, as part of the SHARE 'user statement' concerning the use of data from SHARE, users have to undertake that they will neither make copies of the data available to others nor to enable any third party access to the database and they will take no action aiming at a re-identification of participants. SHARE releases data free of charge to the scientific community, subject to European Union and corresponding national data protection legislation. The data is released in form of so-called 'scientific-use files', which consist of so-called 'factually anonymised' data sets.

In contrast to absolutely anonymised data as contained in 'public-use files', data can be considered as factually anonymised "if they have been altered in such a way that the identity of individuals can only be inferred by expending an unreasonable effort in terms of time, money, and manpower. This type of anonymization is [also] called de facto anonymization" (Rasner, 2012: 62).[40] This concept applied to scientific-use files takes into consideration not only the information content of a data set but also the target group that will be given access to the data, i.e. persons with a scientific affiliation, and aims to apply an adequate degree of

[40] The concept of de facto anonymisation has been elaborated in the context of Section 1, Paragraph 3 of the German Federal Data Protection Act (Bundesdatenschutzgesetz) and in the Social Security Data Protection Act (Sozialdatenschutz) included in the German Social Code (Sozialgesetzbuch), Paragraph 67 of Book X.

anonymisation in relation to their abilities, resources, etc. According to Rasner, the concept of factually anonymised data sets takes account of

> "[t]he high costs of absolute anonymization outweigh its benefits and furthermore, compromise the research value of the data. Anonymization is a trade-off between the risk of personal information being disclosed and the usability of data for research. De facto anonymization makes it almost impossible to re-identify individuals while still providing analytically valid micro-data to researchers." (Rasner, 2012: 63)

As another example of data access and usage restrictions, the Max Planck Language Archive has implemented a system of different levels of access restrictions in order to ensure data confidentiality and to take account of the sensitive nature of certain data (e.g. derived from religious rituals in indigenous communities). The Max Planck Language Archive offers four levels of access:

- "Material under this level is directly accessible via the Internet;
- Material at this level requires that users register and accept the Code of Conduct;
- At this level, access is only granted to users who apply to the responsible researcher (or persons specified by them) and who make their usage intentions explicit;
- Material at this level will be completely closed, except for the researcher and (some or all) members of the speech communities." (Drude et al., 2012: 70)

As the examples from the SSH ESFRI research infrastructures illustrate, there are different ways of dealing with data confidentiality related issues when offering data to wider user groups. Furthermore, different security measures, ranging from anonymisation to data access restrictions and data usage restrictions, or different combinations of such measures can be used to ensure data confidentiality. In this connection, whenever the release of research data is intended, it can already be regarded as a challenging task for researchers to identify the most suitable way of granting access to a certain data set allowing maximising the usability of the data whilst at the same time minimising privacy risks.

Moreover, due to the contemporary fragmentation of data protection laws across Europe it can be a challenging task for cross-country research, to identify which kind of measures can be applied across different countries with regard to access to and re-use of sensitive and/or confidential data. This is amplified by the fact that, depending on national legal provisions, different concepts with regard to released data sets exist – such as the German concept of "scientific use-files", which may only contain 'factually anonymised' data and may only be released to the scientific community. The existence of such national concepts raises the question of whether there are also different notions of what exactly is meant by the terms 'anonymisation' or 'public-use files' (taking into consideration that in Germany, e.g., 'scientific-use files' are defined in contrast to 'public-use files' and therefore both refer to different degrees of anonymisation) when being used in different national or international

contexts. This in turn makes it difficult to identify if (additional) data access and/or usage restrictions have to be employed in different countries when intending to release data sets that are anonymised to a certain degree.

7.3.2 Using and Releasing Paradata (SSc)

In the process of producing survey data much paradata, i.e. micro-level data about the process of survey production[41], are generated. Especially with the increasing use and further development of IT-technologies in survey-based data collection, such as computer-assisted personal interviewing (CAPI) techniques and the implementation of web surveys, the amount of information on the process of survey production has increased.

> "Respondents in web surveys leave electronic traces as they answer survey questions, captured through their keystrokes and mouse clicks. In telephone surveys, automated call scheduling systems record the date and time of every call. In [computer-assisted] face-to-face surveys, interviewers' keystrokes are easily captured alongside the interview and so are audio or even video recordings of the respondent-interviewer interactions. Each of these is an example of paradata available through the computerized survey software." (Kreuter, 2013: 2)

And not only with regard to the collection of paradata a rapid growth can be observed – in the recent years, survey researchers are also increasingly making use of paradata in order "to evaluate and improve survey instruments but also to understand respondents and how they answer surveys" (Couper and Singer, 2013: 57). Furthermore, recently a strong demand from the survey methodology community to make paradata of surveys available can be observed.

Even though "process quality and paradata are not new, a more structured approach in choosing, measuring, and analyzing key process variables is indeed a recent development" (Kreuter, 2013: 2; cf. Couper and Lyberg, 2005). These developments give rise to specific issues which are not covered in existing ethics codes and therefore many legal and ethical issues related to paradata remain unclear.[42]

---

[41] In this report we refer to a broad concept of *'paradata'*, which includes [a] data about the process of survey production recorded as a by-product in the course of conducting a survey (*'process paradata'*), such as listing information, keystrokes, contact data and gross sample data, as well as [b] additional data about the process of survey production obtained separately from external sources or with a specifically targeted effort to enhance the information on the survey production process (*'auxiliary paradata'*), such as interviewer observations, information on the interviewers, external supplementary data about the sample cases, etc. For a more detailed presentation of legal and ethical issues related to the collection and use of paradata please see deliverable D6.2 of the DASISH project (Schmidutz and Bristle, 2013).

[42] It is noted that this currently is a contested area – while some authors claim that the collection and use of paradata is an issue of ethical concern, others argue that the collection and use of paradata does not entail ethical issues at all. In this regard, this report assumes that, if there are claims that the collection and use of

According to Kreuter, especially with regard to releasing paradata "unclear legal and ethical considerations" (Kreuter 2013: 8) can be considered as an obstacle. Up to now, only a few researchers have started to address this issue (Kreuter, 2013: 8) and even these authors state that "[e]xisting ethical codes are not very clear on the issue of paradata" (Couper and Singer, 2013: 58). Moreover, from a legal perspective it is in many cases not clear under which conditions paradata should be collected and how they may be used and be made accessible for re-use to the scientific community.

Since there are several types of paradata that can be collected/recorded in different ways (heavily depending on the way in which a survey is administered) considering the ethical issues and legal requirements that are connected to the collection and use of paradata requires a nuanced approach. For example, with regard to paradata that are unavoidably collected in the process of survey production[43] the only relevant question is whether respondents would consent to their 'use' (cf. Couper and Singer, 2013: 65), while with regard to paradata that are obtained separately from external sources or with a specifically targeted effort[44] the question whether additional[45] consent of the respondents to their collection has to be obtained is of relevance as well.

In all cases, however, there are in particular two issues that are posing legal and ethical challenges to survey researchers. On the one hand, the issue of whether, how and to what extent participants should be informed about the capture and the use of paradata and how much detail should be provided to them. And on the other hand, the issue of how and under which conditions different types of paradata can/may be released for scientific re-use. In relation to both issues, particularly the 'intended use' of the paradata appears to be crucial.

Regarding the first issue, Couper and Singer state:

> "While most […] studies focus on improving the quality of research procedures and, particularly, the questionnaire, paradata are increasingly being used to enhance other information provided by respondents – that is, turning from purely methodological research to more substantive research. There is no consensus on whether, or under what conditions, respondents should be informed that paradata are being collected and may be used. Arguably, they ought to be informed if researchers plan to use such data in conjunction with other information provided by respondents in order to make inferences about individuals. In other words, as the paradata (information about the process) are turned into data (information about respondents), informed consent issues may arise." (Couper and Singer, 2013: 57)

---

paradata is an issue of ethical concern, this subject at least needs ethical consideration. It will be argued that different types of paradata exist and therefore this subject requires a nuanced approach.

[43] I.e. 'process paradata', which are a by-product of survey production.

[44] I.e. 'auxiliary paradata'.

[45] It is supposed, that the respondents have to agree respectively have agreed to participate in the survey.

From an ethical perspective, the crucial question therefore is whether paradata are intended to be used to extend the data sets on the respondents beyond the information provided by them in the course of the survey and under what conditions this may be done. However, "[t]he question of whether the use of paradata […] rises to a level needing explicit mention to respondents" (ibid., 2013: 66) remains difficult to answer.

From a legal perspective, not only European data protection legislation and associated national laws have to be taken into account when trying to assess, under which circumstances and to what extent participants should be informed about the use of paradata, but also – especially if administering web surveys – recent EU online privacy legislation and associated national laws have to considered, such as the "new e-Privacy Directive" (2009/136/EC) (cf. ESOMAR, 2012). In this connection, however, it has to be noted that online privacy laws finally may also affect the collection and use of paradata with regard to survey research in general: "While the intent of [regulations of this kind] is to limit online behavioral tracking, they may encompass a number of more benign activities such as paradata capture in surveys" (Couper and Singer, 2013: 66).

Here, for web-administered surveys with respondents from several countries as well as transnational survey projects, such as SHARE or ESS, the current fragmentation of the European legislative system almost inevitably leads to difficulties when trying to assess the legal requirements for paradata capture and use. However, in this connection, not only national and regional differences in the level of data protection have to be taken into account as long as there is no EU-wide data protection regulation in place; also the nature of the specific kinds of paradata and the mode of collection have to be considered.

However, if one comes to the conclusion that the (intended) use of paradata rises to a level needing explicit mention to respondents, and if it is assumed that

> "respondents are not aware that such additional information is being collected, do not have a reasonable expectation of such capture and use, and, if they were aware of it, might change their behavior or decide not to participate in the survey […], difficult questions arise about how best to provide information about the collection of paradata while at the same time maintaining respondent cooperation with the survey." (Couper and Singer, 2013: 58-59)

This also touches upon the issue of researchers being responsible for ensuring the quality of their research – which according to Singer "is itself increasingly being regarded as an ethical issue" (2008: 96) – since research quality in survey research inter alia depends on the response rates achieved. And, in fact, in several experiments on the effects of asking consent for paradata collection on web survey participation, Couper and Singer (2013: 65) have found that "the concept of paradata is inherently difficult to grasp and is unfamiliar to virtually all respondents [and that t]he potential uses that might be made of such data are equally mysterious [to respondents.]" Furthermore, one major finding of these experiments

was that a change in behaviour actually occurs: "In all three experiments, any mention about capture of paradata lowers stated willingness to participate in a hypothetical survey" (ibid.: 57).

Even though, since the experiments only considered web surveys this finding may not apply with regard to computer-assisted face-to-face or telephone survey data collection[46], these experiments show that the issue of how participants can be informed about paradata collection and use and how much detail should be provided to them while at the same time avoiding a decrease of participation rates remains a challenging task for survey researchers in general.

Considering that there are many different kinds of paradata that can be collected, depending on the survey mode and the technical system in place, and that the various kinds of paradata (such as keystroke data or contact protocols, as collected in the context of SHARE) only can be used for certain kinds of analyses, questions like this one might need to be answered on a case-by-case basis, taking into account the specific kind of paradata, the concrete context in which these data are collected and how they actually are or will be used and released.

With regard to the release of confidential or sensitive paradata, similarly to research data, alternative safeguard measures to anonymisation/pseudonymisation, such as access and usage restrictions, maintaining the usability of the data should be considered when preparing paradata for use and re-use.[47] Additionally, however, the issue of how and under which conditions different types of paradata can/may be released for scientific re-use, is closely related to the question of whether paradata are intended to be used in more substantive research (extending the data sets on the respondents beyond the information provided by them) and under which conditions this may be done.

Since making paradata available to the public or the entire scientific community, would indeed not only make it necessary to consider the 'intended use' but also to consider all ways in which the released paradata possibly could be used – which in turn would impact on the aforementioned consent issues – it appears to be difficult for survey researchers to assess the most appropriate way of releasing certain paradata. Therefore, depending on the nature of the paradata in question, other levels of access providing for special usage restrictions (such as on-site use or remote data access) might be considered as an option.

[46] Paradata capture and use obviously in the experiments have been associated with general threats to privacy occurring on the Internet (e.g. browser-related, IP-related, tracking behaviour of advertisers, hackers and phishers). For example, according to Couper and Singer (2013: 61) many respondents, confused paradata collection with behavioural tracking.

[47] Here, of course, differences in the level of data protection between different EU Member States have to be considered again.

7.3.3 Previously Published (Copyrighted) Language Data (Hum)

In the language sciences, written language data may be obtained from a variety of sources. Many data collections used in linguistics (i.e. text corpora), however, are based on previously published text subject to copyright (e.g. novels, newspapers) and therefore may present challenges related to IPR when they are copied and redistributed for research purposes. Whilst scientific results are to be reproducible, which means that other researchers need to have access to the data, on the one hand, providing such access may constitute a breach of copyright, on the other hand.

Even though some use of copyrighted material is permitted in the USA under the 'fair use limitation'[48] and in Europe through the research exceptions integrated into various legislative regimes, copyright legislation has not kept pace with the current technological developments and the move towards open access. While legislative reform is clearly needed to improve access to language data and to facilitate the replicability of scientific results, licensing of copyrighted language material is increasingly used as an ad hoc and pragmatic solution pending legislative reform. In fact, licensing is becoming an increasingly widespread strategy to deal with the restrictions imposed by copyright legislation, as illustrated by licensing initiatives from the ESFRI communities, such as the CLARIN licensing scheme (see e.g. Gjesdal & Lyse 2013).

An example of the use of licensing schemes for clearing the use of copyrighted text for research purposes are the 'Sofie Analyses'; a language data collection based on copyrighted text[49]. The collection is based on the novel 'Sofies verden' by Jostein Gaarder, and its translation into eight languages (as of July 2013). It takes the form of parallel 'treebanks', i.e. databases of sentences from different languages with detailed information on the grammatical structure. Since the novel 'Sofies verden' has literary as well as linguistic value and since it has been translated into numerous languages, it constitutes a good source for researchers intending to study and compare grammatical patterns across languages. However, since the novel 'Sofies verden' is copyrighted material it would not have been possible to make it accessible to other researchers (users) without negotiating the terms and conditions for scientific re-use with the IPR holders. Moreover, if there are several parties involved in the production of the material, which usually is the case when working with several translations of textual documents, the terms and conditions have to be negotiated with the rights holder of each translation (usually the local publisher or in some cases the translator/s) as well.

Besides text corpora compiled from novels or newspapers, language data sourced from social media may present further ethical and legal challenges with regards to ownership and

[48] In US copyright law, the doctrine of 'fair use' permits limited use of copyrighted material without acquiring permission from the rights holders. Cf. http://www.copyright.gov/fls/fl102.html, accessed 30/06/2013.

[49] See Losnegaard et al. (2013) for a further description of the collection and related work on IPR.

access to the data, as the social media site owner may in fact claim ownership to data produced within the context of their sites, raising complex IPR issues for such data.

Among the social media platforms – from the perspective of the language sciences – Twitter probably is the most interesting and widely used data source, as it offers rich material of everyday language use, which may not only offer insights into new language trends, but also into current political and historical events, such as the Arab Spring, where social media played an important role. Even though some authors claim that tweets in general cannot be considered as copyrightable (see e.g. Reinberg, 2009), Twitter's Terms of Service impose their own limitations on the use of tweets and more specifically on the redistribution of data. Twitter's Developer Rules of the Road[50] state: "If you provide downloadable datasets of Twitter Content or an API [Application Programming Interface] that returns Twitter Content, you may only return IDs (including tweet IDs and user IDs)."[51] As a consequence this means that the number of available tweets in a dataset may fluctuate over time, as e.g. tweets that have been deleted by the user or user accounts that have been set to private will not appear in the 'stream'. While this requirement is certainly understandable in terms of privacy protection and more specifically with regard to the "right to be forgotten"[52], this may pose problems if the data is used in scientific research; especially with regard to the replicability of scientific results if the data set is constantly changing due to removal of 'old' material.

8 Concluding remarks

8.1 Legal & Ethical Issues in the Social Sciences and the Humanities

As the previous chapters have shown, researchers in the social sciences and the humanities currently are facing many ethical and legal challenges, some of which are similar, such as those related to the legal and ethical framework. Correspondingly, when new technologies

[50] Available at https://dev.twitter.com/terms/api-terms (last update: July 2, 2013).

[51] This has been further specified by a representative of the Twitter API Policy in the Twitter Developers' Forum: "Under our API Terms of Service (https://dev.twitter.com/terms/api-terms), you may not resyndicate or share Twitter content, including datasets of Tweet text and follow relationships. You may, however, share datasets of Twitter object IDs, like a Tweet ID or a user ID. These can be turned back into Twitter content using the statuses/show and users/lookup API methods, respectively. You may also share derivative data, such as the number of Tweets with a positive sentiment. As such, if you would like to share this data set, you will need to remove the tweet text and creation date from the data set and replace these with the appropriate Tweet ID." (Source: https://dev.twitter.com/discussions/3021, accessed 30/06/2013)

[52] Besides, the use of Twitter data also raises other privacy issues that have to be considered. While it is not certain that Twitter usernames can be used to identify individuals, the dissemination of usernames also may affect users' privacy on other levels, as argued by Petrović et al. (2010). In this perspective, in order to avoid "malicious use of the data (e.g., by spammers)" (ibid.: 1), the creators of the Edinburgh Twitter Corpus decided to replace original Twitter usernames with an ID.

are employed over the course of research data generation, management and dissemination an amplification of ethical issues can be observed. Furthermore, legal and ethical challenges, such as obtaining consent from data subjects (i.e. respondents, participants, informants) by some means or another occur in relation to all SSH ESFRI research infrastructures that are collecting data for scientific research purposes. However, even though such common general challenges can be identified this does not necessarily mean that all concrete aspects subsumed under these general topics are of relevance for both the social sciences and the humanities domain or even all of the five SSH ESFRI research projects. As the examples ("use cases") presented in Section 7 illustrate, there are commonalities, but there are also differences in relation to the legal and ethical challenges.

In general, challenges related to ethics issues and legal requirements that concern the day-to-day operations of SSH data collection, curation and dissemination, can be experienced on tree different levels:

(1) Regarding the legal and administrative framework;
(2) With regard to general ethics issues and legal requirements;
(3) Related to specific issues.

Besides the aforementioned commonalities that can be located on the first two levels and which are of a rather general nature, with regard to concrete and special issues and practical solutions, due to intrinsic differences in substance and methodology[53] between research in the humanities and the social sciences, many differences on the third level become apparent as well. While, on the one hand, for example, obtaining informed consent from third party individuals usually is not a major issue for population-based surveys in the social sciences, ensuring the autonomy of third parties involved in humanities data collection frequently poses a problem. On the other hand, informed consent issues related to the capture of paradata, which specifically relates to survey research, does not seem to present an issue for humanities research at the present stage. Furthermore, the concrete ways of obtaining informed consent from data subjects – in particular with regard to the questions of whether, how and to what extent participants should be informed – heavily depend on the type of data being collected as well as on the concrete research context in which these data are collected and how they are to be disseminated and may be used.

Particularly in the context of WP6 of the DASISH project – which includes a very broad range of SSH research projects from the collection and analyses of text and speech corpora through to transnational socio-economic survey research – it becomes obvious that a pragmatic view about the commonalities as well as the differences between social sciences and humanities research has to be developed when approaching concrete legal and ethical challenges. In WP6, on the one hand some general common interests could be identified,

[53] E.g. with regard to qualitative and quantitative methodologies or concerning the ways of research data generation (collection of primary data vs. use of secondary data), etc.

such as to understand and evaluate the possible effects that the anticipated European General Data Protection Regulation may/will have with regard to data collection and long-run data preservation in the SSH domain. Especially with regard to the legislative regime which impacts upon the governance of the research process, common challenges could be identified.

On the other hand it became increasingly clear that concrete legal and ethical issues in many cases only concern a few of the involved research infrastructures and that particularly in connection with the actual day-to-day operations of the SSH ESFRI research infrastructures specific challenges occur that consequently call for specific solutions and in many cases require a nuanced approach (e.g. legal and ethical issues related to paradata; cf. Section 7.3.2). For example, legal and ethical challenges related to the linking of administrative data to survey data are of common interest of CESSDA and SHARE. ESS and SHARE, for instance, have a clear common interest in legal and ethical issues related to the use and dissemination of paradata and with regard to the issue of data ownership in the context of transnational survey research. In contrast, e.g. CLARIN is concerned with legal and ethical challenges related to recordings of spoken language data or IPR issues connected to the (re)distribution of previously published written documents for research purposes. These issues, which also are of importance for DARIAH, usually do not occur in relation to population-based survey research in the social sciences infrastructure projects.

Being able to differentiate between issues and needs that concern several (sometimes all, many times only a few) of the SSH ESFRI research infrastructures participating in the DASISH project and those which do not, is not only of crucial importance with regard to the cooperative work in the DASISH project[54] but also – and this is more important – with regard to the development of guidelines for appropriate data protection measures or standards for procedures requiring legal and ethical consideration (such as data linkage procedures, et cetera). Efforts of advising and guiding researchers and RIs in order to support them in coping with legal and ethical challenges they experience as a result of contemporary data collection, integration, linking and sharing practices in their respective fields of research can only be successful if being tailored to specific subjects/topics and to the needs of those which actually are affected by issues related to these subjects/topics.

8.2 Ethical Guidelines and Guidance for Researchers

As noted in Section 3.2, numerous codes of ethics and associated guidance on good professional conduct and research integrity are available to the research community (cf. Denscombe, 2002; Table in Section 3.2, Annex 10.2). New technologies and practices, recent

[54] Taking into consiceration the main objective of the DASISH project; i.e., to provide the five ESFRI research infrastructure projects with common solutions to common issues and challenges.

developments have given rise to, amplify rather than transform ethics in social sciences and humanities research. Codes of ethics, however, rarely give concrete guidance in the face of ethical dilemmas; their purpose should be to alert researchers to the need to consider the consequences of their research activities. The International Statistical Institute's Declaration on Professional Ethics, for example, presents a clear statement of the function of the Declaration – it does not attempt to resolve difficult ethical choices:

> "Instead it offers a framework within which the conscientious statistician should be able to work comfortably. It is urged that departures from the framework of principles be the result of deliberation rather than of ignorance." (ISI Declaration on Professional Ethics, 2010: 3)

A feature of research ethics is that they are based on values; the determination of the correctness of a decision is open to interpretation. Ethics codes are not legally binding, being contingent on context. Researchers' responses to ethics dilemmas can be informed by pragmatic as well as other considerations. However, the fundamental nature of legal requirements relating to data protection, privacy, data sharing and so on is that they are binding and their infringement will incur sanctions. Accordingly, the RESPECT Code of Practice for Socio-Economic Research is based on a synthesis of a number of existing professional and ethical codes of practice, together with current legal requirements in the EU. It states:

> "Whilst the RESPECT provisions are voluntary, some of the requirements on which they are based are morally binding on the members of specific professional associations or legally binding on citizens of EU Member States." (RESPECT, 2004)

The key principles of research ethics have been reviewed above (see Section 3), relating to informed consent and confidentiality, within the broad dictum 'do no harm'. The application of existing codes of ethics, designed to guide researchers, may not address specific ethics issues arising in research using information and communication technology nor inter-disciplinary SSH research which uses data collection methods from other disciplines.

With regard to issues related to e-social science, for example, Charlesworth has pointed out that the degree to which ethics guidelines are *directly helpful* in addressing these issues not only heavily depends on the interests of the groups who have developed them but also on the extent to which they have been subject to continuing review and revision (cf. Charlesworth, 2012: 93).

The issues of informed consent et cetera have to be re-examined in the light of recent technological developments, e.g. social media, an increased interest in data mining and big data research. Lomborg notes that "a key issue in questions in current debates over Internet ethics is whether such research endeavours involve human subjects or not" (2012: 21) – this is a non-trivial question, raising issues such as 'perceived privacy' and digital identity. As he

comments, "[t]he way internet phenomena, and with this, the data of internet researchers are conceptualised in regard to personhood will determine whether the research involves human subjects or not" (2012: 22). Further, whether information is deemed 'sensitive' or 'non-sensitive' – a legal category – means that it may or may not come under the remit/purview of national data protection agencies (Lomborg cites the Danish example, 2013: 25) and be assessed in accordance with national arrangements.

One element of the governance of science relates to ethics review; in many countries, academic institutions have established committees with responsibility for overseeing the ethical dimensions of research projects. As noted in Section 4, often the requirement for research ethics is linked to institutional liability requirements.

Especially for cross-national surveys application for approval of ethics committees may constitute a major challenge, particularly if activities that fall into the health research domain are included (e.g. regarding the collection of DBS in SHARE). For SSH research projects with cross-disciplinary approaches that, for example, include the collection of biological samples, and therefore currently require the involvement of bio-medical research ethics committees in many countries, the present fragmentation of the national ethics committee systems in Europe constitutes a serious problem. Since in such cases approval has to be given by several ethics committees on national and in several European countries even on a regional/local level (e.g. Switzerland, Italy, Belgium), even identifying all responsible ethics committees appears to be challenging.

Furthermore, the need for bio-medical ethics boards to appreciate the SSH ethics requirements and to appreciate the differences between clinical and social science research is a challenge reported by SSH researchers. Therefore, a major challenge for surveys within the SSH domain is the harmonisation of ethics committee approval procedures across European countries.

The RESPECT code seeks

> "not to create new requirements or restrictions on the conduct of research, but to protect researchers from unprofessional or unethical demands and to raise awareness of ethical issues and spread existing professional good practice, enabling the development of a European Research Area with common standards that are transparent and universally agreed. Such common standards are a prerequisite for the development of a European market in socio-economic research, in which research can be commissioned and partnerships entered into on the basis of clear mutual understandings and expectations." (RESPECT, 2004)

With the Clinical Trials Directive, the European Union (EU) envisioned a harmonisation of research ethics committees (RECs) across Europe, a similar proposal for SSH research has not been advanced yet. The EC Report "Global Governance of Science – Report of the Expert Group on Global Governance of Science" (2009) notes that

> "Although the European ethical consensus may be more or less accepted by many countries, its enactment varies widely. The UNESCO Declaration, too, allows for a variety of implementations even though the wording is universal. In practice, global declarations, attempting to harmonise ethical standards, often end up at the lowest common denominator. Even so, resulting values may be prioritized differently in different regions, cultures and traditions. There may be no such thing as a set of 'European' ethical values, but there are clearly tensions between European and some other approaches to ethics, such as those more typical of the United States. In the United States, for example, there is a tendency for *autonomy* to outweigh *dignity* in ethical decision making, whereas the opposite is the case in Europe." (European Commission, 2009: 30)

The Expert Group concludes

> "The challenge therefore is to encourage the harmonisation of ethical values as part of a long-term project of global reflection on ethics, while recognizing and learning from diverse ethical practices." (ibid.)

This only can be achieved through on-going dialogue between key stakeholders.

8.3 Legal Framework and the Data Protection Reform

Law and legal practice in relation to various types of data do affect opportunities for SSH research as well as the possibilities for data archives and SSH research infrastructures to serve the needs of empirical research. Therefore, all SSH ESFRI research infrastructures should devote special attention to developments this area, and in particular to the current reform of the European data protection legislation.[55]

At present, the "Data Protection Directive" (95/46/EC) regulates the protection of individuals with regard to the processing of personal data and the free movement of such data. Since the Directive has failed to achieve proper harmonisation of data protection laws across the European countries, currently the fragmentation of data protection law in the EU Member States poses a big challenge to cross-country SSH research and hampers the free exchange of personal data between European countries.

Due to technological change and particularly the growing importance of the Internet (social media, cloud computing, etc.) the European Commission proposed a new regulation in January 2012, updating the existing legislation. The proposed General Data Protection

[55] Please also see http://dasish.eu/publications/presentations/ for a presentation with regard to the current development of the proposed General Data Protection Regulation and its possible implications on research data collection, data preservation and data sharing in the SSH domain that has been held as part of the DASISH IASSIST session in May 2013 in Cologne, Germany (Kvalheim, 2013).

Regulation is intended to replace Directive 95/46/EC and associated national/regional data protection legislation. As a legal instrument that is directly applicable the anticipated Regulation aims to harmonise the legal practice and ensure a unified legal data protection framework in Europe.

Certainly due to the mere fact of reducing[56] fragmentation of data protection law across Europe, the anticipated Regulation is expected to simplify many procedures in transnational SSH research, especially regarding the transfer of data across national borders.

However, due to on-going controversial discussions and negotiations there is a lot of uncertainty with regard to the provisions of the Regulation. European researchers are concerned about the changes the General Data Protection Regulation might bring about and how these will affect research data collection, processing and dissemination in the future. At this, for the SSH research community the question is whether the new Regulation will provide good, safe and predictable conditions for research.

Even though the current fragmentation of data protection law in Europe constitutes a major problem for cross-national research, the current legal instrument has its positive aspects as well. Most importantly, the Directive provides an exemption from the 'purpose limitation principle', which can be considered as a fundamental research guarantee, in particular for register based research (including the linkage of survey data with administrative record data). According to the Directive further processing of personal data for historical, statistical or scientific purposes is not considered as incompatible with the original purposes, and therefore may be performed (provided that appropriate safeguards are taken).[57] Furthermore, personal data may be stored for longer periods of time than necessary for the purposes for which the data were collected in case of historical, statistical or scientific use (if the public interest clearly exceeds the disadvantages).[58]

The comprehensive reform of the data protection rules as initially proposed by the Commission also for the most part accommodates research interests and implies more continuity than change in conditions. Although in the provision on the purpose limitation principle in Article 5b of the Regulation[59] the clarification that further processing of personal data for scientific research is not incompatible with the original purpose has been dropped,

[56] In this regard, it is noted that – even though the proposed Regulation "will do away with many complexities and inconsistencies stemming from the different implementing laws of the Member States currently in place" (EDPS, 2012: 4) – in the proposed Regulation still quite a lot of space for coexistence and interaction between EU law and national law remains. In some cases provisions of the Regulation clearly built on national law respectively allow/mandate national law to build on and in other cases provisions allow/require national law to specify/further develop rules in certain areas or even to depart from the provisions (cf. ibid.: 9). However, in the Committees' proposal the provision of Article 83, which addresses the "processing for historical, statistical and scientific research purposes", does not allow for specific national rules (cf. ibid.: 49).

[57] Cf. 95/46/EC, Article 6, Paragraphs 1b.

[58] Cf. 95/46/EC, Article 6, Paragraphs 1e.

[59] Article 5b of the Regulation corresponds to Article 6, Paragraphs 1b in Directive 95/46/EC.

Recital 40 to some extent repair this, stating that: "The processing of personal data for other purposes should be […] allowed […] in particular where the processing is necessary for historical, statistical or scientific research purposes." Furthermore, the Commission's proposal in Article 6 lists alternative grounds for lawful processing of personal data, which mainly means continuity in the conditions for processing personal data for scientific purposes. In addition it even provides a 'new' Paragraph 2, which explicitly authorises the processing of personal data for research purposes: "Processing of personal data which is necessary for the purposes of historical, statistical or scientific research shall be lawful subject to the conditions and safeguards referred to in Article 83."

On the whole, the Commission's proposal for a General Data Protection Regulation can be summarised as follows:

- The core data protection principle of purpose limitation is strengthened
- The rights of the data subject are strengthened: consent (explicit), information, the right to be forgotten
- Information, knowledge and consent, the most important measures to safeguard privacy are strengthened
- The role and responsibilities of the data controller (institution) is strengthened
- The designation of a data protection officer is mandatory[60] and will be the main element in the system for regulating, controlling and documenting the processing of personal data

Altogether, the proposal can be understood as an effort to increase the level of data protection. Nevertheless, it can be said that the Commission's proposal is striking the right balance between the public interest in information privacy and research, not least because of the new 'research provision' including Article 83 and its associated provisions containing research exemptions/guarantees and protecting the public interest in research.

Regarding this, however, several amendments that have been proposed recently, especially those put forward in the "Draft report on the proposal for a regulation of the European Parliament and of the Council on the protection of individual[s] with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)" of the Committee on Civil Liberties, Justice and Home Affairs (16.01.2013) – henceforth "the Albrecht-Report" – signal a shift of balance.

The suggested amendments of the Albrecht Report to the Commission's proposal for a General Data Protection Regulation have caused widespread and serious concern in research environments across Europe. On the one hand, the Albrecht-Report supports the Commission's aims of strengthening the rights of the data subject, ensuring a unified legal framework and reducing the administrative burdens for the data controller. On the other hand, however, the Albrecht Report suggests dropping more or less all the important

[60] According to Article 35 of the proposal this applies to all public authorities and public bodies as well as to enterprises with more than 250 employees.

research provisions (derogations) that grant research a privileged position with regard to access and use of personal data. It argues that scientific research is not special with regard to its public interest, and hence does not deserve a privileged position within the legal framework.

Accordingly, In Amendment 27 (Proposal for a regulation, Recital 42) concerning derogations from the prohibition on processing sensitive categories of data, the Albrecht-Report argues that

> "[p]rocessing of sensitive data for historical, statistical and scientific research purposes is not as urgent or compelling as public health or social protection. Consequently, there is no need to introduce an exception which would put them on the same level as the other listed justifications." (Albrecht-Report, 2013: 24)

Regarding Recital 50 of the Commission's proposal, which concerns exemptions from the duty to inform the data subject, the Albrecht-Report keeps the provision that "it is not necessary to impose this obligation where the data subject already disposes of this information, or where the recording or disclosure of the data is expressly laid down by law, or where the provision of information to the data subject proves impossible or would involve disproportionate efforts" (Proposal for a General Data Protection Regulation 2012: 25), but deletes the following passage:

> "The latter could be particularly the case where processing is for historical, statistical or scientific research purposes; in this regard, the number of data subjects, the age of the data, and any compensatory measures adopted may be taken into consideration" (Albrecht-Report, 2013: 26-27),

arguing that, "[t]he deleted text may be misunderstood as promoting a lower level of protection for certain kinds of data processing (ibid.: 27). Furthermore, in Amendment 327 concerning Article 81 of the Commission' Proposal on the processing of personal health data, the processing of personal data concerning health which is necessary for historical, statistic or scientific research purposes is limited and "shall be permitted only with the consent of the data subject" (Albrecht-Report, 2013: 197-198), arguing that "health data is extremely sensitive and deserves utmost protection" (ibid: 198). According to Amendment 328, "health data, which is extremely sensitive, may only be used without the consent of the data subject if it serves an exceptionally high public interest and [if it is] anonymised or at least pseudonymised using the highest technical standards." (ibid: 198-199).

Finally, with regard to the Amendments 334-337, which concern the article on "Processing for historical, statistical and scientific research purposes" (Article 83) of the proposal of the European Commission, the Albrecht-Report states that

> "[i]n cases where the data subjects have not given consent, sensitive data and data about children should only be used for research purposes if based on law

and serving exceptionally high public interest. Otherwise, any 'research', no matter if academic or corporate and including e.g. market research, could be used as an excuse to override all protections provided for in the other parts of this Regulation" (Albrecht-Report, 2013: 203).

According to these amendments, as a rule, data about children as well as sensitive data[61] can only be used for (any kind of) research if consent of the data subject has been obtained. Member States may only provide exemptions on condition that the research serves an exceptionally high public interest and in this case only if the data are anonymised, or, if this is not possible, at least pseudonymised.

On the whole, for the scientific research in all SSH fields and particularly for register based research, including linking survey data with administrative record data, these amendments are devastating. By removing many important research provisions/derogations granting research a privileged position with regard to data access and the use of personal data, the Albrecht-Report is clearly contradicting high level policies for open access and data sharing across Europe.

Due to the possible negative consequences for SSH research, which the anticipated General Data Protection Directive may have, from an SSH perspective, the further development of the proposed Regulation and its possible effects concerning the collection, processing and dissemination of different types of data occurring in the SSH domains should be closely observed. Furthermore, since the European institutions are currently entering a crucial stage in the legislative process, whenever possible the opportunity should be taken to call the attention of research funding institutions and ministries (among others) to the damaging effects that the proposed amendments of the Albrecht-Report will have on research and society when being implemented. From SSH perspective – fully aware of researchers' ethical responsibilities and legal obligations – an adequate legal framework has to be developed to safeguard both privacy and autonomy and access to personal data for scientific purposes.

8.4 Present and Future Ethical & Legal Challenges of SSH Research

In summary, besides core elements of research ethics – such as 'do no harm', informed consent, protection of anonymity and confidentiality – which not only govern ethical considerations of researchers conducting research involving human subjects but also are crucial with regard to ethics committee approvals, law relating to, for example, data

[61] I.e. according to the Albrecht-Report: "personal data, revealing race or ethnic origin, political opinions, religion or philosophical beliefs, sexual orientation or gender identity, trade-union membership and activities, and the processing of genetic data or data concerning health or sex life or criminal convictions, or related security measures" (Albrecht-Report, 2013: 80).

protection and copyright and database rights can be identified as legal provisions that are particularly relevant for the conduct of research.

Concerning contemporary or current research in the SSH domains, various legal and ethical issues can be identified that SSH researchers and SSH RIs are confronted with in the course of data collection, data processing and data curation. For each stage of the research process, general ethical and legal issue (such as obtaining informed consent, anonymisation/ pseudonymisation, data access and usage restriction) can be identified. Besides, specific challenges connected to different types of data and/or different ways of data collection have to be considered as well. These specific issues are strongly connected to recent technological developments and the cross-national nature of many research projects in the European Union; they amplify the nature of research ethics and give rise to specific challenges at different stages of the research process, such as:

- Obtaining additional informed consent related to the capture and use of different types of paradata;
- Ethics committee approval when collecting biological samples in the context of population-based transnational survey research;
- Ensuring privacy and autonomy of third party individuals when collecting recordings of spontaneous speech;
- Data protection measures to be taken, when processing and storing biological samples in population based survey research;
- Safeguarding of personal data and confidential information at every stage of data linkage (e.g. when linking survey data with administrative record data), including the challenge of anonymisation/pseudonymisation of indirect identifiers and/or relational data;
- IPR issues related to the collection and harvesting of previously published language data from the Internet and other copyrighted language data.

In general, these challenges heavily depend on the type of data being collected as well as on the concrete research context in which these data are collected and how they are to be disseminated and may be used. However, certain general legal and ethical issues, which concern all SSH ESFRI RIs in some way or another, in particular with regard to the ethical, legal and administrative frameworks in Europe, can be identified that are of particular relevance with regard to present and future research in the SSH domains:

- Technological developments and in particular the Internet not only open doors to new and enriching possibilities in research but also pose a challenge to researchers, since they amplify ethical challenges and give rise to specific issues which are not covered in existing ethics codes.
- The organisation of the national ethics committee systems in Europe differs a lot between the different EU Member States, and the same applies for the

approval procedures. When conducting transnational and/or transregional research projects that require approval of various national/regional ethics committees (e.g. cross-disciplinary research including the collection of biological samples), identifying all committees responsible and applying to them in accordance with the respective national or regional policies and procedures may pose serious problems and even may prevent research.

- Since the Data Protection Directive (95/46/EC), which currently regulates the protection of individuals with regard to the processing of personal data within the European Union, has failed to achieve proper harmonisation of data protection laws across the European countries, currently the fragmentation of data protection law in the EU Member States poses a big challenge to cross-country SSH research.
- On-going controversial discussions and negotiations about the provisions of the proposed new European General Data Protection Regulation, however, causes uncertainties regarding the implications of the Regulation with regard to research data generation and management in the SSH domains and the extent to which the Regulation will affect the work of the existing data archives and SSH research infrastructures. Certainly some of the recently proposed amendments would have serious negative consequences for SSH research (and in particular for register based research) if being transposed into EU legislation.

Since some of the specific legal and ethical challenges are related to or amplified by these general legal and ethical problems, with regard to future research it is important not only to solve the problems related to the very specific challenges that research in the SSH domain is confronted with but also to tackle the issues that are of a rather general nature. However, since the specific challenges are most pressuring from the perspective of the RIs as well as individual researchers, as these are part of the day-to-day operations of SSH data collection, efforts to find solutions should consider these first.

In this connection, of course, in relation to ethics, the challenges have to be addressed on a case-by-case basis; for those of a legal nature, breach of legal requirements is generally more clearly identified. With regard to future legal challenges, obviously the most significant issue raised at present relates to the forthcoming European General Data Protection Regulation – since the provisions of this Regulation will provide the legal framework for future research in the SSH domains. However, responsible research that involves human subjects also – taking into account the amplification of ethics issues due to the use of new technologies and practices in SSH research, nowadays perhaps more than ever before – to consider the (traditional) ethical dimensions of 'informed consent' and 'confidentiality protection' and 'justice' have to be considered (cf. Singer, 2008: 80).

9 References

Adolphs, S., Knight, D. and Carter, R. (2011). Capturing context for heterogeneous corpus analysis. Some first steps. International Journal of Corpus Linguistics 16 (3).

Alcser, K., Antoun, C., Bowers, A. , Clemens, J. and Lien, C. (2012). Ethical Considerations in Surveys Guidelines. In: CSDI Cross-Cultural Survey Research Guidelines, Section III. Retrieved June 29, 2013, from http://www.cessda.org/sharing/rights/3/.

Article 29 Data Protection Working Party (2007). Opinion 4/2007 on the concept of personal data. 01248/07/EN, WP 136, Adopted on 20th June. Retrieved June 26, 2013, from http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf.

Bainbridge, D. (2007). Intellectual Property. Pearson Education Ltd: Harlow.

Brundell, P., Tennent, P., Greenhalgh, C., Knight, D., Crabtree, A., O'Malley, C., Ainsworth, S., Clarke, D., Carter, R. and Adolphs, S. (2008). Digital Replay system (DRS): A Tool for Interaction Analysis. Proceedings of the 2008 International Conference on Learning Sciences (Workshop on Interaction Analysis): Utrecht.

Bulmer, M. (2001). The Ethics of Social Research. In Gilbert, N (ed.). Researching Social Life, Sage: London, pp. 45 -57.

Bulmer, M. and Warwick , D. (1993). Social Research in Developing Countries: Surveys and Censuses in the Third World. UCL Press: London.

Calderwood, L. and Lessof, C. (2009). Enhancing Longitudinal Surveys by Linking to Administrative Data. In: Lynn, P. (ed.). Methodology of Longitudinal Surveys). John Wiley & Sons, Ltd: Chichester, UK, pp. 55-72.

CESSDA (n.d.). Data Protection & Confidentiality. In: Council of European Social Science Data Archives - CESSDA. Retrieved June 27, 2013, from http://www.cessda.org/sharing/rights/2/ and http://www.cessda.org/sharing/rights/3/.

Charlesworth, C. (2012). Data protection, freedom of information and ethical review committees: Policies, practicalities and dilemmas. Information, Communication & Society, Vol. 15 (1), pp. 85-103.

Couper, M. P. and Lyberg, L. (2005). The Use of Paradata in Survey Research. Proceedings of the 55th Session of the International Statistical Institute.

Couper, M. P. and Singer, E. (2013). Informed Consent for Web Paradata Use. Survey Research Methods 7 (1), pp. 57-67.

Denscombe, M. (2002). Ground Rules for Good Research: A 10 point guide for social researchers. Open University Press: Buckingham.

Drude, S., Broeder, D., Wittenburg, P. and Sloetjes, H. (2012). Best practices in the design, creation and dissemination of speech corpora at The Language Archive. Proceedings from the workshop: Best Practices for Speech Corpora in Linguistic Research. Retrieved June 25, 2013, from http://www.corpora.uni-hamburg.de/lrec2012/Proceedings_Complete.pdf.

EDPS – European Data Protection Supervisor (2012): Opinion of the European Data Protection Supervisor on the data protection reform package. 7 March 2012. Retrieved June 20, 2013, from http://www.europarl.europa.eu/document/activities/cont/201205/20120524ATT45776/20120524ATT45776EN.pdf.

EDPS – European Data Protection Supervisor (2013): Additional EDPS Comments on the Data Protection Reform Package. 15 March 2013. Retrieved June 25, 2013, from https://secure.edps.europa.eu/EDPSWEB/webdav/site/mySite/shared/Documents/Consultation/Comments/2013/13-03-15_Comments_dp_package_EN.pdf.

ESOMAR (2012). "ESOMAR Practical Guide on Cookies." Retrieved June 27, 2013, from http://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ESOMAR-Practical-Guide-on-Cookies_July-2012.pdf.

European Commission (2009). Global Governance of Science – Report of the Expert Group on Global Governance of Science to the Science, Economy and Society Directorate, Directorate-General for Research, European Commission. Publications Office: Luxembourg. Retrieved June 30, 2013, from http://ec.europa.eu/research/science-society/document_library/pdf_06/global-governance-020609_en.pdf.

European Commission (2012a). Areas of untapped potential for the development of the European Research Area (ERA): Analysis of the response to the ERA Framework public consultation. Publications Office, Luxembourg. Retrieved June 30, 2013, from http://ec.europa.eu/research/era/pdf/analysis-of-response-era-consultation_en.pdf.

European Commission (2012b). Ethical and Regulatory Challenges to Science and Research Policy at the Global Level. Publications Office: Luxembourg. Retrieved June 30, 2013, from http://ec.europa.eu/research/science-society/document_library/pdf_06/ethical-and-regulatory-challenges-042012_en.pdf.

Flew, A. (ed.) (1979). A Dictionary of Philosophy. Pan, London, pp. 104-105.

Fulton, J. (2012). Respondent Consent to Use Administrative Data. College Park, MD: University of Maryland, Joint Program in Survey Methodology, PhD dissertation.

Gill, L. (2001). Methods for Automatic Record Matching and Linkage and their Use in National Statistics. National Statistics Methodology Series, No. 25.Her Majesty's Stationery Office: London. Retrieved June 26, 2013, from http://www.ons.gov.uk/ons/guide-method/method-quality/specific/gss-methodology-series/index.html.

Gjesdal, A. M. and Lyse, G. I. (2013). CLARIN licensing schemes. Retrieved June 20, 2013, from http://dasish.eu/dasishevents/iassistws/.

Iphofen, R. (2011). Ethical decision-making in social research. A practical guide. Palgrave Macmillan: London.

Kimmel, A. J. (1988). Ethics and values in applied social research. Applied Social Research Methods Series, Vol.12, Sage: London.

Korbmacher, J. and Czaplicki, C. (2013). Linking SHARE Survey Data with Administrative Records: First Experiences from SHARE Germany. In: Malter, F. / Börsch-Supan, A. (eds.). SHARE Wave 4: Innovations & Methodology. MEA, Max Planck Institute for Social Law and Social Policy: Munich, pp. 47-52.

Kreuter, F. (2013). Improving Surveys with Paradata: Analytic Uses of Process Information. Wiley.

Kvalheim, V. (2013). New Legal Challenges – New EC Privacy Regulation, Data Preservation and Data Sharing in danger? Presentations at the DASISH IASSIST session, May 30 2013, Cologne, Germany. Available at http://dasish.eu/publications/presentations/.

Lomborg, S. (2013). Personal internet archives and ethics. In: Research Ethics, 9(1), pp. 20-31.

Losnegaard, G. S. et al. (2013). Linking Northern European Infrastructures for Improving the Accessibility and Documentation of Complex Resources. In: Proceedings of the workshop on Nordic language research infrastructure at NODALIDA, May 22-24, 2013, Oslo. NEALT Proceedings Series 20. Retrieved June 30, 2013, from: http://www.ep.liu.se/ecp/089/005/ecp1389005.pdf.

Oyen, E. (ed.) (1990). Comparative Methodology. Theory and Practice in International Social Research. Sage Studies in International Sociology 40.

Parry, O. and Mauthner, N. (2004). Whose data are they anyway? Practical, legal and ethical issues in archiving qualitative research data. In: Sociology, 38(1), pp. 139-152.

Petrović, S., Osborne, M. and Lavrenko, V. (2010). The Edinburgh Twitter corpus. In Workshop on Social Media, NAACL 2010. Retrieved June 30, 2013, from: http://homepages.inf.ed.ac.uk/miles/papers/socmed10.pdf.

Rasner, A. (2012). The distribution of pension wealth and the process of pension building – Augmenting survey data with administrative pension records by statistical matching. Dissertation, Technische Universität, Berlin. Retrieved June 30, 2013, from: http://opus.kobv.de/tuberlin/volltexte/2012/3384/.

RESPECT (2004). The RESPECT Code of Practice. In: Respect Project - Professional and Ethical Codes for Socio-economic Research in the Information Society. Retrieved June 28, 2013, from http://www.respectproject.org/code/.

Robson, C. (1993). Real World Research. Blackwell: Oxford.

Ryan, L., Cooper, P. and Drey, N. (2013). University Research Ethics Committees as learning communities: identifying training needs to support effective decision making. Research Ethics Vol.9.

Reinberg, C. (2009). Are Tweets Copyright-Protected? In: WIPO MAGAZINE. Retrieved June 30, 2013, from http://www.wipo.int/wipo_magazine/en/2009/04/article_0005.html.

Sampson, G. (2000). CHRISTINE Corpus, stage i: Documentation. Technical report, Sussex: University of Sussex.

Schmidutz, D. and Bristle, J. (2013). Sample Merged Paradata Sets: Ethical and Legal Issues. DASISH Deliverable D6.2. Available at http://dasish.eu/deliverables/.

Singer, E. (2008). Ethical Issues in Surveys. In: De Leeuw, Edith D. / Hox, Joop J. / Dillman, Don A. (eds.). International Handbook of Survey Methodology. Psychology Press, Taylor & Francis: New York, pp. 78-96.

Stenström, A.-B., Andersen, G. and Hasund, K. (2002). Trends in Teenage Talk. Corpus compilation, analysis and findings. John Benjamins Publishing Company: Philadelphia.

Swatman, P. (2012). Ethical issues in social networking research. Seminar presentation.

UK Data Archive (n.d.). Anonymisation. In: Create and Manage Data. Retrieved June 27, 2013, from http://www.data-archive.ac.uk/create-manage/consent-ethics/anonymisation

UKAN (n.d.). What is Anonymisation? In: UK Anonymisation Network. Retrieved June 27, 2013, from http://www.ukanon.net/key-information/.

UK Data Forum (2009). UK Strategy for Data Resources for Social and Economic Research, 2009 – 2012, National Data Strategy.

UNECE – United Nations Economic Commission for Europe (2009). Principles and Guidelines on Confidentiality Aspects of Data Integration Undertaken for Statistical or Related Research Purposes. United Nations. Retrieved June 30, 2013, from http://www.unece.org/fileadmin/DAM/stats/publications/Confidentiality_aspects_data_integration.pdf.

Van Wel, L. and Royakkers, L. (2004). Ethical issues in web data mining. Ethics and Information Technology, pp. 129-140.

World Medical Association (2008). WMA Declaration of Helsinki - Ethical Principles for Medical Research Involving Human Subjects. In: Website of the World Medical Association. Retrieved June 29, 2013, from http://www.wma.net/en/30publications/10policies/b3/index.html.

10 Annex

10.1 Acronyms and Abbreviations

CESSDA – Council of European Social Science Data Archives

CLARIN – Common Language Resources and Technology Infrastructure

CORDIS – Community Research and Development Information Service (EC)

DASISH – Data Service Infrastructure for the Social Sciences and Humanities

DoW – Description of Work, Annex 1 to the Grant Agreement of the DASISH project

DARIAH – Digital Research Infrastructure for the Arts and Humanities

L&E – Legal and Ethical: L = Legal / E = Ethical

EC – European Commission

EDPS – European Data Protection Supervisor

EU – European Union

ERIC – European Research Infrastructure Consortium

ESFRI – European Strategy Forum on Research Infrastructures

ESRC – The Economic and Social Research Council (UK)

ESS – European Social Survey

FP7 – Seventh Framework Programme for Research and Technological Development (EC)

IPR – Intellectual Property Rights: IP = Intellectual Property

RECs – Research Ethics Committees

RIs – Research Infrastructures

RDA – Remote Data Access

RDC – Research Data Centre

SHARE – The Survey of Health, Ageing and Retirement in Europe

SSH – Social Sciences and Humanities: SSc = Social Sciences / Hum = Humanities

WP(#) – Work Package(Number)

10.2 Selection of Ethical Guidelines and Codes of Ethics

10.2.1 General Principles for the Treatment of Subjects

- **Helsinki Declaration** (originally adopted by the World Medical Assembly in 1964, sixth revision 2008): WMA Declaration of Helsinki – Ethical Principles for Medical Research Involving Human Subjects
- **Belmont Report** (Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, United States, 1979, created under the National Research Act of 1974): Belmont Report: Ethical Principles and Guidelines for the Protection of Human Subjects of Research

10.2.2 Codes of Ethics for Survey Professionals

- **WAPOR Code of Ethics** (revised WAPOR Code of Ethics, effective 1 December 2011): Code of ethics of the World Association for Public Opinion Research
- **AAPOR Code of Ethics** (code of ethics of the American Association for Public Opinion Research, revised May 2010): AAPOR Code of Professional Ethics and Practices
- **ICC/ESOMAR Code of Ethics** (code of ethics of the European Society for Opinion and Market Research, first promulgated in 1948, revisions include provisions of the International Chamber of Commerce, 1994): ICC/ESOMAR International Code on Marketing and Social Research Practice

10.2.3 Codes of Ethics of Different Scientific Disciplines

10.2.3.1 Sociology & Economics

- **RESPECT Code of Practice** (voluntary code covering the conduct of socio-economic research in Europe, RESPECT project, 2004): RESPECT Code of Practice for Socio-Economic Research (please see: An EU Code of Ethics for Socio-Economic Research)
- **Code of Ethics of the American Sociological Association** (approved by the ASA Membership in June 1997): ASA Code of Ethics and Policies and Procedures of the ASA Committee on Professional Ethics
- **ISA Code of Ethics** (approved by the ISA Executive Committee, Fall 2001): International Sociological Association's Code of Ethics
- **Code of Ethics of the British Sociological Association** (BSA, March 2002, including appendix "Further sources of information, advice and support", updated in May 2004): Statement of Ethical Practice for the British Sociological Association
- **GSE & DGS Code of Ethics** (code of ethics of the German Sociological Association (GSE) and the Berufsverband Deutscher Soziologen (BDS), 27 November 1992, in German language only) Ethik-Kodex der Deutschen Gesellschaft für Soziologie (DGS) und des Berufsverbandes Deutscher Soziologinnen und Soziologen

10.2.3.2 Psychology

- **APA Code of Ethics** (2003, with the 2010 amendments): Ethical Principles of Psychologists and Code of Conduct of the American Psychological Association
- **Ethical guidelines of the German Association of Psychology** (including the Code of Conduct of the Association of German Professional Psychologists, April 1999): Ethical Principles of the German Psychological Society (DGP) and the Association of German Professional Psychologists
- **Codes of Ethics of National Psychology Organisations** (Source: International Union of Psychological Science – Psychology Resources Around the World)

10.2.3.3 Biology & Medicine

- **CIOMS/WHO Ethical Guidelines** (2002): International Ethical Guidelines for Biomedical Research involving Human Subjects, prepared by the Council for International Organizations of Medical Sciences (CIOMS) in collaboration with the World Health Organization

10.2.3.4 Statistics

- **Declaration on Ethics of the International Statistical Institute** (adopted by the ISI Council 22 & 23 July 2010): ISI Declaration on Professional Ethics
- **Code of Ethics of the American Statistical Association** (prepared by the Committee on Professional Ethics, approved by the Board of Directors, August 7, 1999): Ethical Guidelines for Statistical Practice
- **Fundamental Principles of Official Statistics of the United Nations Statistics Division** (the United Nations Statistical Commission, in its Special Session of 11-15 April 1994, adopted the Fundamental Principles of Official Statistics, earlier set out in the Economic Commission for Europe's Decision C (47), but incorporating a revised preamble): Fundamental Principles of Official Statistics

10.2.3.5 Linguistics

- **Max Planck Institute DOBES project (Documentation of Endangered Languages)**: Documents on Ethical and Legal Aspects
- **Sources for Thinking About the Ethics of Sociolinguistic Research**: An overview of useful sources

10.2.4 Further Links/Sources

- **RESPECT project links on Ethical Codes and Guidelines** (Source: RESPECT project, 2004, Institute for Employment Studies)
- Association of Internet Researchers' (AoIR) Ethics Guide

10.3 Extract from the 'RESPECT Code of Practice'

(Source: http://www.respectproject.org/code/clegal.php?id=, accessed 28/06/2013)

2. Compliance with the law

In general, socio-economic researchers should comply with the laws of the countries in which they are based or in which they are carrying out research. In the case of international collaborations or online research, the laws of additional countries may also apply. Researchers have a duty to ensure that their work complies with any relevant legislation. Two areas of law (data protection law and intellectual property law) are particularly relevant for the conduct of research, especially research involving human subjects, and researchers should acquaint themselves with the relevant national and international provisions.

2.1 Data protection

2.1.1 Legal requirements

Socio-economic research often involves the collection and other further processing of personal data. The processing of personal data is regulated by law, and researchers have therefore to comply with the relevant national legislation of the current Member States of the European Union that implement the European Directive 95/46/CE.

In order to comply with the terms of the data protection law, researchers should:

a. find out whether the processing will include personal data (ie, not just confidential data but any data related to an identifiable individual)
b. examine which national law applies, especially in international co-operations
c. determine who will be the person responsible for the processing (the controller)
d. collect the data only for specified, explicit and legitimate purposes
e. collect only data that are adequate, relevant and not excessive with regard to the purpose of the processing
f. keep the data accurate and, where necessary, keep them up-to-date
g. process the data fairly and lawfully
h. in general, not keep data longer than necessary according to the purpose of the processing and when the purpose is achieved, destroy or render the data anonymous. In some countries where personal data may be kept for longer periods for historical, statistical or scientific use, researchers may keep them longer if all the conditions for this longer storage are fulfilled.
i. not further process the data in a way incompatible with the initial purpose(s). If the data are further processed for scientific or statistical purposes, researchers should comply with requirements regarding the re-use of personal data
j. respect the conditions regarding the legitimacy of the processing, bearing in mind that to qualify as legitimate it must meet one of the social justifications laid down by the law
k. comply with the information duty towards data subjects to provide information on the identity, address of the controller, purpose of the processing, and other information stipulated by law unless an exemption is provided by the law
l. comply with duties towards National Data Protection Authorities by providing the required information regarding the planned processing and, where relevant, obtaining prior consent, unless an exemption is provided by the law

m. respect the rights of data subjects to access personal data, rectify incomplete or inaccurate data, and to object to the processing under the stipulated circumstances
n. take technical and organisational measures to ensure the security and confidentiality of personal data (including encryption where necessary)
o. comply with the conditions for communication of personal data to third parties or recipients, bearing in mind that it is only lawful to transfer data if the purpose is compatible with that for which the data were originally collected
p. refrain from transferring personal data outside the European Economic Area except where an adequate level of protection has been acknowledged by the European Commission or if not, except if the legal conditions provided by the relevant law are respected.

2.1.2 Good practice

Good practice, as embodied in existing professional codes, lays out the following principles, which aim at ensuring the security and confidentiality of personal data.

a. Researchers in socio-economic studies are obliged to protect personal data, ie information on identifiable individuals. In order to prevent misuse of data, data are to be stored properly and adequately (eg, by storing information through which individuals can be identified, separately from the remaining research material). Particular caution is necessary in this context with regard to the risks posed by electronic data processing and data transfer.
b. Researchers should respect the anonymity, privacy and confidentiality of individuals participating in the research, and ensure that the presentation of data and findings does not allow the identity of individuals participating in a study, or informants, to be disclosed or inferred. Researchers should also ensure that this is also the case in the presentation of findings by contractors, funding agencies or colleagues. In cases where disclosure of the identity of a subject (whether an individual or an organisation) is central and relevant to the research such confidentiality cannot always be guaranteed. In such cases the problem should be addressed in open discussion with research subjects, with the aim of obtaining informed consent to any disclosure.

The security and confidentiality of data is only one aspect of data protection; the other legal requirements are still compulsory. Therefore, research should be conducted in accordance with all the principles of the applicable national data protection legislation.

Before embarking on the collection of any personal data, researchers should take into account the duties and conditions of processing, make an analysis of the processing envisaged, identify the operations that will be involved and the level of sensitivity of the data, in order to assess the lawfulness of the exercise.

2.2 Intellectual property

European directives on intellectual property converge with professional good practice in requiring researchers to pay attention to ensuring necessary permissions, correct attribution of authorship, acknowledgement of sources, correctness of references and the avoidance of plagiarism.

2.2.1 Legal requirements

Wherever practicable, intellectual property rights should be explicitly addressed in contracts covering the conduct of socio-economic research, whether these are funding contracts, partnership agreements or employment contracts.

In accordance with European directives and national legislation on intellectual property rights, the following questions and principles should be taken into account when conducting socio-economic research:

a. recognising the relevance of intellectual property rights to socio-economic research
b. taking due account of the fact that (especially in an online environment and/or international co-operations) several national laws might be applicable that differ substantially from the regulations in the researcher's home country
c. paying due respect to the fact that material used in socio-economic research is predominantly protected by intellectual property rights such as copyright, database and software protection
d. ascertaining which acts within typical research conduct are unacceptable without (statutory or contractual) permission due to rights being reserved for the author under intellectual property legislation (as named above)
e. realising how exceptions/exemptions/limitations supersede individual permission for certain acts of socio-economic research under certain conditions
f. understanding how to use licences and assignments of rights when creating or using material protected as intellectual property
g. taking into account how employment contracts might affect intellectual property
h. realising the consequences of copyright infringements.

In order to comply with intellectual property law, socio-economic researchers should:

a. find out to what extent questions of intellectual property rights (copyright, database and software protection) are concerned in the particular research activity
b. examine which countries' laws apply, especially in international co-operations and when using the Internet
c. assume that any material created or used in socio-economic research might be intellectual property and consider protection before using it
d. realise that many ways of using protected material – such as reproduction by down-/upload or by paper/digital copies, publication, making material available on the Internet, alteration (eg, for online format etc.) – are generally reserved for a rightsholder, and find out when permission is therefore (in principle) required
e. when relying on legal permission (like the exceptions for quotation, research or 'fair use') for any particular conduct, consider carefully the respective extent and conditions
f. if a planned activity is not clearly covered by statutory permissions (for example quotation rights) identify the rightsholder and conclude authorising contracts (transfer/assignment of rights/license agreements). Ascertain that the permission covers explicitly all relevant aspects – among them the description of type, extent, duration, environment (such as online) of the intended use, any preparatory or subsequent acts, rights involved, responsibility for possible infringements, remuneration etc.
g. where several parties are involved (researchers, assistants, funding parties, employment situations in institutes, enterprises, universities) ensure explicit consensus among parties in advance, about rights matching the intended use.

2.2.2 Good practice

Good practice in relation to intellectual property goes beyond the bare legal requirements. Existing professional codes lay out the following principles:

a. In principle, authorship is reserved for those researchers who have made a significant intellectual contribution to a research project, the writing of a research report or another scholarly piece of work. Seniority and position in a research institution's hierarchy alone is not sufficient for authorship. Honorary authorship is unacceptable. In cases where several persons collaborate on a research project or publication, the question of authorship and intended use of the results should be discussed, and consensus achieved among participating researchers as early on in the project as possible. The order of authors listed should take account of their respective contributions to the work. All collaborating researchers, whether named as authors of a publication or not, bear responsibility for the contents of the respective publications and the presentation of data and findings in these publications.
b. Any third parties' material protected by copyright must be clearly identified and clearly attributable to their original authors, regardless of the form their presentation and quotation might take (except in cases where it is necessary for the original author to remain anonymous; in such instances, however, it must be made clear that the information was provided by an anonymous person). Lack of permission for a given use is considered as theft of intellectual property. Even if material, including data, sources, information or ideas drawn from the work of others is not protected by copyright, it should be identified as third parties' material. Failure to acknowledge the original authorship of such material, as well as knowingly presenting ideas, methodologies and research findings of others in ways that may lead observers to suppose that they are one's own, is regarded as plagiarism and is unacceptable.

2.3 Other laws

A wide range of other laws may also apply, varying from general health and safety, employment and anti-discrimination laws, to specific regulations governing the appointment and management of researchers, and more specific regulations that may govern the context in which particular kinds of research are carried out.

There may be certain circumstances that form exceptions to this rule, for instance when criminal behaviour itself forms the subject of the research undertaken. In such cases, researchers should:

- raise the matter with research funders
- ensure that full documentation is maintained to establish the bona fide nature of the research, and
- where necessary, seek the advice of their relevant professional association.

In more extreme cases, research may be carried out in countries where democratic government is absent, or relatively recent, and certain laws are considered to be inherently unjust, socially harmful or detrimental to scientific integrity. In such cases too, individual researchers must take responsibility for decisions of professional judgement and their professional associations have a responsibility to support them.

10.4 Transnational Inquiry on Ethics Committee Approvals

As part of WP6 of DASISH a transnational systematic inquiry regarding national legal requirements and ethics committee approval procedures with regard to the collection of biomarkers (derived from dried blood spots) has been carried out by MEA (MPG) making use of the SHARE Research-Network. The following set of questions was sent to the persons responsible for the application for ethics committee approval in the SHARE Country Teams:

A. Ethics committee approval procedure

I. Responsibility and contact

- How does the ethics committee system of your country work in terms of scopes? Is it organised on an institutional or on a regional or local level?
- Which ethics committee or IRB (Institutional Review Board) is responsible for the approval procedure for the DBS collection in your country?
- Is a single ethics committee responsible for the approval of the DBS collection for the whole country or are there several ethics committees that have to be consulted?
- Who is/are the contact person/s of the ethics committee/s within your country?

II. Steps

- Which steps have to be taken in order to 'get the DBS through' the ethics committee?
- How will our (resp. your) request/application be processed by the ethics committee/s internally (from the submission of the request via the review by the committee/s to the delivery of an approval certificate)?

III. Documents and materials

- What materials and documents are needed by the ethics committee/s in order to be able to review the DBS collection?
- At what stage (of A.II.) must the DBS materials/documents be handed in?
- Are there any other requirements to be met? Is further information requested (e.g. about the SHARE survey as a whole)?

IV. Duration

- How often (and when) does/do the responsible ethics committee/s meet?
- Are there any specific dates that have to be taken into account?
- Approximately, how long does an approval procedure take?
- If we are asked by your ethics committee/s to implement changes, will the procedure take considerably more time in this case?

V. Costs

- Are there any fees or other foreseeable costs connected to the approval procedure of your ethics committee/s?
- Approximately, how much costs can be expected to 'get the DBS through' the approval procedure necessary in your country?

VI. Other important issues

- Are there any other important issues that should be considered when applying for an ethics committee approval within your country?

B. Legal framework

I. Collection of the DBS

- Are there any important legal constraints that have to be taken into account when collection DBS (DBS specific or regarding biological material in general) in your country? E.g.:
  - Are (trained) laypersons allowed to prick the fingers of the participants in order to collect the DBS?

II. Type of consent

- Depending on contemporary national legislation, (informed) consent sometimes has to be given in a written form, sometimes verbal consent is sufficient. What type of consent has to be given in your country?
- Are there any special requirements regarding the 'amount' of information which has to be given to the participants prior to their consent and the way in which they have to be informed?

III. Data privacy and consent forms

- May the DBS samples be sent to another country for analyses (i.e. to Denmark in our case)?
- May consent forms (containing personal information) be sent to another country in general? In other words, would we be allowed to send the consent forms to the laboratory?
- May both, the DBS consent form and the blood sample be stored (separately) at the laboratory at the same time in general?

IV. Liabilities

- Is it necessary to take out a specific "study participant insurance" for all persons participating in the DBS collection or will the "public liability insurance" of your institution cover harm (which is unlikely to occur, by the way) that might result from the DBS collection?

V. Other requirements/restrictions

- Are there any other important legal requirements or restrictions that should be taken into account when planning to collect DBS in your country?

C. Ethics committee/s – probable requirements/restrictions

I. Collection of the DBS

- Do we have to expect any specific or general reservations about the intended collection of DBS of (one of) "your" ethics committee/s? E.g.:
  - Might the ethics committee/s demand – extending beyond "your" country-specific legal constraints – that only medical staff will prick the fingers of the participants in order to collect the DBS?
  - Might the ethics committee/s demand that you take out a "study participant insurance" (see B.III.)?

II. Transmission of DBS analyses results

- Is it rather likely that the ethics committee/s of your country will commit us to inform the participants or their general practitioners about the blood results in case a value lies outside the normal range – or is it maybe even more likely that they advise us not to do so?

III. Other foreseeable requirements/restrictions

- Are there any other requirements or restrictions (which go beyond legal constraints) that can be expected to be considered to be important by "your" ethics committee/s that might make adaptions of our DBS procedures and/or materials necessary?